TEXTE ET CORPUS :

Actes des Troisièmes Journées de la Linguistique de Corpus

TEXTE ET CORPUS :

Actes des Troisièmes Journées de la Linguistique de Corpus

Sous la direction de Geoffrey Williams

SOMMAIRE

Sommaire 5

Présentation des auteurs 9

Introduction
Geoffrey Williams 11

Lexicométrie et corpus 17

Où l'on mesure la distance entre les distances
Étienne Brunet 19

Analyse lexicométrique de l'opposition générique dans une perspective endogène
Margareta Kastberg Sjöblom 41

Le logiciel Hyperbase : préparation et présentation des corpus Application à un corpus de langue anglaise
Vasilica Milea 61

Approche lexicométrique de corpus d'interactions verbales entre un adulte et un enfant en cours d'acquisition du langage. Résultats d'expérience
Luiggi Sansonetti 71

Linguistique et corpus 87

Néologie du portugais brésilien
Ieda Maria Alves 89

Observations linguistiques sur un corpus de légendes en anglais de photographies de trains
Pierre J. L. Arnaud, François Maniez 99

Le corpus et la pragmatique : une hypothèse sur l'emploi contrastif de l'imparfait et du passé composé
Lidia Fraczak, Stéphanie Giron 113

Opposition entre *de* et *des* devant les noms précédés d'épithète en français : portée du « poids »
Itsuko Fujimura, Mitsumi Uchida, Hiroshi Nakao 131

Stratégie de consultation de corpus oraux transcrits : pistes méthodologiques pour l'exploration d'un corpus thématique à haut rendement
Nathalie Gasiglia 145

Le discours direct dans le corpus comparable spécialisé
Aurelija Leonavičienė 165

Utilisation d'un corpus catégorisé pour l'étude et la représentation de la synonymie en contexte
Jean-Luc Manguin 181

BDVOX : Base de Données pour Systèmes de Reconnaissance de la Parole Multilocuteur
I. C. Seara, F. S. Pacheco, R. Seara Jr., S. G. Kafka, S. Klein, R. Seara 197

Corpus, vous avez dit corpus ! De la notion de corpus à la création d'un « corpus informatisé »
Céline Vaguer 207

TAL et corpus 225

Qu'est-ce qu'un « corpus homogène » ? Réflexions à partir d'expériences en Extraction et Recherche d'Information
Patrice Enjalbert 227

Appariement de mots : propagation des liens d'équivalence à l'aide de la relation syntaxique Sujet
Sylwia Ozdowska 239

Repérage des non-phrases thématiques grâce à la plate-forme ContextO
Sylvie Porhiel.....251

Apport de l'analyse linguistique pour l'extraction terminologique en corpus : application au domaine de la génomique
Fabienne Ville-Ometz, Alain Zasadzinski, Dominique Besagni 269

Les besoins d'interactions en traitement automatique des langues et en linguistique de corpus : étude de cas
Stéphane Ferrari, Vincent Perlerin.....289

PRÉSENTATION DES AUTEURS

Ieda Maria ALVES	Université de São Paulo (Brésil)
Pierre J. L. ARNAUD	Université Lumière-Lyon II, CRTT
Dominique BESAGNI	Unité de Recherche et Innovation (URI), INIST − CNRS UPS76
Étienne BRUNET	Université de Nice
Patrice ENJALBERT	Université de Caen − GREYC, CNRS
Stéphane FERRARI	Université de Caen − GREYC, CNRS UMR 6072
Lidia FRACZAK	Université de Clermont-Ferrand II, LRL
Itsuko FUJIMURA	Université de Nagoya (Japon)
Nathalie GASIGLIA	Université de Lille III − SILEX, CNRS UMR 8528
Stéphanie GIRON	Université de Clermont-Ferrand II, LRL
S. G. KAFKA	Université Fédérale de Santa Catarina (Brésil), LINSE
Margareta KASTBERG SJŐBLOM	ILF − CNRS Bases, Corpus et Langage UMR 6039
S. KLEIN	Université Fédérale de Santa Catarina (Brésil), LINSE

Aurelija LEONAVIČIENĖ	Universitas Vytauti Magni (Lituanie)
Jean-Luc MANGUIN	Université de Caen − CRISCO, CNRS UMR 6170
François MANIEZ	Université Lumière-Lyon II, CRTT
Vasilica MILEA	Université de Metz
Hiroshi NAKAO	Université d'Aïchi (Japon)
Sylwia OZDOWSKA	Université de Toulouse le Mirail
F. S. PACHECO	Université Fédérale de Santa Catarina (Brésil), LINSE
Vincent PERLERIN	Université de Caen − GREYC, CNRS UMR 6072
Sylvie PORHIEL	Université de Chypre, Nicosie, Université Paris IV − LaLLIC, CNRS UMR 8139
Luiggi SANSONETTI	Université Paris III − EA 2290 SYLED − EA 170I CALIPSO, ILPGA
I. C. SEARA	Université Fédérale de Santa Catarina (Brésil), LINSE
R. SEARA	Université Fédérale de Santa Catarina (Brésil), LINSE
Mitsumi UCHIDA	Université féminine d'Osaka (Japon)
Céline VAGUER	Université Paris X − MoDyCo, UMR 7114
Fabienne VILLE-OMETZ	Unité de Recherche et Innovation (URI), INIST − CNRS UPS76

INTRODUCTION

Geoffrey Williams
Université de Bretagne Sud

Corpus et texte, texte et corpus. Quels sont les liens entre ces deux termes ?

En réalité la réponse est compliquée par le choix même des termes corpus et texte. Tout le monde sait que 'texte' n'est pas un concept simple. Est-ce qu'un chapitre, ou un article de presse, doit être considéré comme une unité à part entière, ou par rapport à un ensemble qui est lui-même le texte ? Est-ce qu'un simple énoncé peut être vu comme texte ? La situation se complique dans le cas de l'oralité. Est-ce que le texte se limite uniquement à des discours préalablement écrits, et qui, par conséquent, forment un ensemble cohérent avec un début et une fin clairement définis, ou est-ce que des conversations avec des voix multiples peuvent aussi être considérées comme texte ? C'est un peu comme le mot, 'mot', nous acceptons ce que nous voulons dire, et l'ambiguïté inhérente. Le texte est forcément transcrit, dépasse de préférence l'énoncé, et est authentique. Nous sommes obligés d'accepter que les textes prennent des formes très variables selon les contextes d'emploi. C'est pour cette raison que dans sa célèbre définition du terme 'corpus', Sinclair (1996) a préféré parler de « pieces of language ».

La définition du terme 'corpus' est compliquée par une variété d'emplois et des différences importantes dans les cultures de recherche en langues entre les mondes anglo-saxon et français. Nous pouvons dégager quatre grandes familles de corpus :

- en littérature - un recueil d'œuvres, souvent d'un même auteur.
- en linguistique de corpus – une collection de textes sélectionnés par le linguiste selon des critères précis afin de représenter une langue.
- en linguistique énonciative – une collection d'énoncés authentiques, sélectionnés par le linguiste pour les besoins de son étude.
- En linguistique générative – un ensemble de phrases non-authentiques, créées par le linguiste pour les besoins de ses études.

Il me semble évident que pour la linguistique de corpus, il faut écarter les deux derniers types de corpus. Leur but est en effet largement syntaxique puisque le contexte est trop restreint pour analyser un sens textuel. L'étude

des inter-relations entre la syntaxe et le lexique est impossible sans le contexte au sens large. L'outil de base en linguistique de corpus est le concordancier, mais il n'est efficace que sur des textes de taille suffisante pour mériter une analyse informatisée.

C'est avec les deux autres types de corpus, littéraires et corpus constitués, que les traditions linguistiques peuvent provoquer une confusion. En France le prestige de la littérature française et des études littéraires est omniprésent. La grande archive textuelle FRANTEXT est largement littéraire et les deux grands dictionnaires, *Larousse* et *Robert*, utilisent des citations tirées des grandes œuvres. En ce qui concerne le lexique, une certaine tradition prescriptive reste de mise. De l'autre côté de la Manche, l'approche est plus pragmatique. La tradition littéraire reste importante dans les grands dictionnaires comme le *Oxford English Dictionary*, mais depuis la guerre, la mondialisation de la langue anglaise a créé une tradition de dictionnaires pour apprenants où la phraséologie et les exemples tirés de contextes non littéraires sont de mise. Les dictionnaires comme le *Oxford Advanced Learner's Dictionary* dominent le marché international. La révolution opérée par le dictionnaire *COBUILD* dans les années 80 n'était pas seulement une révolution lexicographique avec le premier dictionnaire basé sur corpus, mais également une révolution dans le monde de la linguistique outre-Manche. La linguistique de corpus, basée sur des corpus constitués selon des critères précis (Atkins et al. 1992, Biber 1993) est devenue la norme.

Les deux types de corpus, et donc d'approches, sont-ils incompatibles ? La réponse est évidemment non. L'optique ne sera pas la même, mais les outils peuvent l'être. Le but diffère d'une approche à l'autre : un corpus littéraire est surtout employé pour caractériser le style d'un auteur ou d'un genre, et le corpus constitué pour caractériser et généraliser sur une langue. Il suffit de respecter les spécificités et les exigences des deux approches.

Au fond, qu'est ce qu'un corpus ? Une collection de textes sous format numérique ? La nature du corpus prototypique variera selon l'approche, mais dans les deux cas, le corpus est constitué par un assemblage de textes homogènes et assemblés dans un but précis. Il est même impératif de ne pas oublier le texte puisque notre forêt linguistique est constituée d'arbres textuels et la forêt ne doit pas cacher les arbres qui la constituent. De plus, nous ne pouvons pas véritablement parler de corpus, mais la même approche peut parfaitement bien être appliquée à un seul texte si la contrainte de taille suffisante est respectée. L'important est de pouvoir comparer des textes afin de voir autrement.

Les Journées de la Linguistique de Corpus 2003 ont cherché à mettre à l'honneur les deux types de corpus et les deux approches aux textes que sont la linguistique de corpus *per se* et la lexicométrie. Il n'y a pas de rupture

entre les deux, mais un continuum dans lequel nous pouvons trouver d'autres disciplines de la linguistique qui utilisent des corpus. Il faut ajouter également des études qui sont plus tournées vers l'analyse informatique. Afin de montrer la grande variété des études sur corpus nous avons choisi de diviser ces actes en trois sections, lexicométrie et corpus, linguistique et corpus et TAL et corpus. Il va de soi que ces coupures ne sont pas parfaites, la catégorisation simple ne fonctionne jamais dans les domaines pluridisciplinaires.

Les travaux d'Etienne Brunet illustrent parfaitement les objectifs de la lexicométrie littéraire. Son texte traite de deux controverses : la relation entre Corneille et Molière, et la méthodologie adoptée par Labbé. Il s'agit donc d'une analyse littéraire et d'une discussion approfondie des positions méthodologiques.

L'article de Margareta Kastberg-Sjoblom s'attaque à la notion de genre littéraire vue à travers une analyse factorielle. Cependant, au lieu d'étudier une gamme d'auteur, elle analyse l'œuvre d'un seul, Le Clézio, afin d'analyser l'évolution de genres dans ses écrits.

Le point commun entre les trois premiers textes de la section Lexicométrie et Corpus est l'outil Hyperbase, créé à l'origine pour l'analyse de textes en français mais repris par la suite par d'autres chercheurs travaillant surtout sur des langues romanes. Dans le cas des études de Vasilica Milea il s'agit d'un corpus littéraire anglais, 67 nouvelles d'Edgar Allen Poe. Son article constitue une analyse des fonctions statistiques de Hyberbase et des contraintes pour la préparation d'un corpus en anglais.

Le quatrième article dans cette section, celui de Luigi Sansonetti, applique la lexicométrie à l'analyse d'un corpus oral. Il s'agit d'une transcription de trois dialogues entre un adulte et un enfant. Lexico 3 est employé afin regarder le discours des trois enfants, puis l'analyse factorielle permet de comparer leur discours et celui de l'adulte.

La section Linguistique et Corpus est beaucoup plus variée. Nous n'entrons pas dans le débat sur la différence entre la linguistique de corpus, où le corpus est central et où la méthodologie est à la base d'une discipline indépendante, et la linguistique sur corpus, où différentes disciplines de la linguistique appliquée ou formelle ont recours à un corpus numérique. L'important ici est la variété des approches et des domaines étudiés.

La lexicographie et le développement de la lexicographie moderne vont de pair. Ainsi, l'article d'Alves décrit une recherche de néologismes en Portugais Brésilien. L'axe de cette recherche est une analyse morphologique sur le Brésilien, mais la méthodologie est adaptable à d'autres langues, et la tâche importante pour le lexicographe.

La communication de Arnaud et Maniez met en exergue la notion de texte dans un corpus. Dans un corpus de légendes, le texte est réduit, mais en combinant les méthodologies de la linguistique de corpus et l'analyse de genre, ils mettent en évidence un nano-genre souvent ignoré puisque dans la constitution de corpus les légendes sont souvent écartées en même temps que les images qu'elles décrivent.

Il se peut que la France ait pris un certain retard sur les corpus écrits, mais ceci est loin d'être le cas pour l'oralité. L'article de Fraczak et Giron est une étude sur l'emploi de l'imparfait et du passé composé avec des répercussions importantes dans l'enseignement du Français Langue Etrangère, le FLE.

L'écart entre règles et réalité se fait aussi entendre dans la communication de Fujimura, Uchida et Nakao. Dans ce cas, l'étude porte sur un corpus constitué d'articles de presse, de forums de discussions, de débats parlementaires et de FRANTEXT. L'analyse porte aussi bien sur la question diachronique que sur le genre textuel.

Gasiglia nous ramène à l'oral, mais dans un domaine particulier : le football. L'article traite de la méthodologie de consultation sur un corpus enrichi utilisant le balisage en XML pour faciliter des requêtes.

Jusqu'ici les corpus ont été monolingues ; l'étude de Leonaviciene, par contre, est une étude comparative du français et du lithuanien. L'analyse de discours direct dans les textes politiques montre les stratagèmes discursifs employés, et également l'influence de l'oral sur l'écrit.

Il est bien connu que la synonymie totale est quasi inexistante, mais la synonymie partielle est omniprésente. Les travaux de Manguin traitent de la synonymie dans FRANTEXT en utilisant comme base de départ le dictionnaire de synonymes en ligne du CRISCO, de l'Université de Caen. Un tel travail peut servir non seulement à enrichir le dictionnaire, mais également à améliorer des systèmes de désambiguïsation.

Avec l'article de Seara et al., nous passons du texte écrit à l'oral dans l'objectif de la reconnaissance de la parole. Bien que l'objectif soit la reconnaissance, le texte ne décrit pas seulement les applications, mais surtout les travaux de constitution et d'analyse de corpus préalables à toute mise en œuvre de système**.** Le texte est une analyse approfondie des problèmes à résoudre dans la constitution d'un corpus et le développement d'une base de données multilangues.

Le dernier texte de cette section remet en cause la définition même du corpus. La France a une longue tradition de grammaire d'énonciation où domine l'étude de phrases authentiques. Dans l'article de Vaguer, nous avons un corpus énonciatif basé sur un corpus de textes avec des phrases

sélectionnées, puis traitées avec des outils bureautiques classiques. Nous ne sommes pas dans la linguistique de corpus *per se*, mais dans une autre façon de voir un corpus numérisé.

La dernière section est consacrée au Traitement Automatique de la Langue sur corpus. Il faut souligner une fois de plus que cette catégorisation est arbitraire puisqu'il y a continuité entre les approches. Dans cette section nous avons surtout des applications informatisées de la recherche sur corpus, dans des contextes très variés, ce qui montre le dynamisme du domaine.

Pour être représentatif, un corpus doit être homogène, soit par la thématique, soit, pour un corpus de référence, par la langue décrite, même si l'homogénéité ne va pas de soi. Enjalbert étudie cet aspect vital de la constitution des corpus du point de vue de l'extraction et de la recherche d'information. L'explosion des ressources électroniques disponibles rend facile l'assemblage d'un « corpus », mais l'homogénéité est loin d'être garantie. La question posée par Enjalbert, et sa réponse, sont extrêmement pertinentes pour le développement de la discipline et les applications basées sur corpus.

Pendant trop longtemps, les linguistes ont séparé le lexique de la syntaxe alors que les recherches sur corpus ont clairement montré que les deux sont indissociables. Les liens ont été clairement montrés dans les grammaires locales (Gross 1994) et Pattern Grammars (Hunston 2000). Le défi pour l'analyse des comportements des mots en contexte est de dépasser les annotations morphosyntaxiques pour réaliser des analyses fonctionnelles. L'article de Ozdowska va dans ce sens avec une étude utilisant l'analyseur SYNTEX. L'objectif est ici l'appariement de mots, tâche essentielle en traduction.

L'article de Porhiel nous ramène à l'écrit avec la description d'un outil pour le repérage des prétendues non-phrases thématiques. Toutes les phrases ne contiennent pas nécessairement un verbe, ce que les correcteurs automatiques ont du mal à admettre. Ces phrases existent et peuvent être porteuses d'informations, il faut donc les étudier à travers un corpus afin de réaliser la richesse des formes et de trouver les formalismes pour leur traitement.

La terminologie représente un autre domaine important en linguistique de corpus. En TALN, l'analyse des corpus permet l'extraction des termes *in situ*. Ville-Ometz et al. se situent résolument dans le domaine de l'ingénierie linguistique avec la description d'une méthodologie pour la reconnaissance des termes et la prise en compte des variations. La plate-forme décrite introduit des filtres linguistiques dans les métarègles afin d'améliorer l'extraction. Les avantages et les limites d'une telle approche sont analysés.

Le dernier article dans ce recueil est celui de Perlerin et Ferrari. L'article constitue une conclusion tout à fait pertinente à ces Actes qui illustrent la richesse des approches possibles en linguistique de corpus, puisqu'il s'agit d'explorer la nécessité d'interaction entre le TAL et la linguistique de corpus. La situation est complexe quand il s'agit de présenter des outils TAL à des novices qui ne sont pas conscients des contraintes posées par la constitution d'un corpus. Dans cette introduction au modèle LUCIA et à l'outil LUCIABUILDER, une manière de combler cette lacune est décrite avec une étude de cas.

Il peut y avoir un risque de fracture entre les pratiquants d'un TAL formalisant où le corpus n'existe que pour valider des résultats obtenus par la machine et une linguistique descriptive qui utilise les corpus pour comprendre et décrire l'interaction humaine. Cependant, chaque discipline peut apporter à l'autre si des ponts et des lieux d'échange existent.

Les Journées de la Linguistique de Corpus sont un lieu de rencontres : rencontre entre disciplines, rencontre entre approches et surtout rencontre entre chercheurs. L'ambiance conviviale, la confrontation d'idées sont des aspects cruciaux. Les textes assemblés dans ce recueil montrent le dynamisme des études autour des corpus et servent de témoignage à ce dynamisme, comme apport à la linguistique en générale, et comme base de départ pour d'autres débats lors des prochaines journées.

Geoffrey Williams
Université de Bretagne Sud
Lorient

REFERENCES

Atkins, B.T.S, Clear, J, Ostler, N. 1992. "Corpus Design Criteria". *Literary and Linguistic Computing. Journal of the Association for Literary and Linguistic Computing*. 7/1 : 1-16

Biber D. 1993. "Representativeness in Corpus Design". *Literary and Linguistic Computing*. *Journal of the Association for Literary and Linguistic Computing*. 8/4 243-257.

Gross M,. 1994. "Constructing lexicon grammars" in ATKINS, B.T.S. and ZAMPOLLI, A. 1994. Computational Approaches to the Lexicon. Clarendon Press: Oxford.

Hunston, S., Francis, G. 2000. *Pattern Grammar: A corpus-driven approach to the Lexical Grammar of English.* Amsterdam and Philadelphia: Benjamin's

Sinclair J. 1996. *Preliminary recommandations on Corpus Typology*, EAG--TCWG--CTYP/P. Version May 1996. Pisa : EAGLES. http://www.ilc.cnr.it/EAGLES96/typology/typology.html (consulté le 16/02/2006)

LEXICOMETRIE ET CORPUS

OÙ L'ON MESURE LA DISTANCE ENTRE LES DISTANCES

Étienne Brunet
Université de Nice

1. INTRODUCTION

L'exposé qui va suivre reprend et prolonge le thème d'une conférence prononcée en avril à la Sorbonne dans le cadre d'un cycle intitulé, non sans humour, « Tous ceux qui comptent ». Un journaliste, qui se trouvait dans la salle, s'est fait l'écho, plus sonore que fidèle, des propos que j'ai cru devoir tenir dans l'affaire Corneille-Molière qui s'étalait alors sur la place publique, et où mon nom avait été imprudemment cité. L'article, paru le 11 avril 2003 dans l'hebdomadaire *Le Point*, passait sous silence les longs développements que j'avais consacrés à la méthode prônée par Dominique Labbé, pour n'en retenir que la conclusion, laquelle contestait l'interprétation donnée aux faits observés, mais non pas leur mesure. L'affaire s'est envenimée dans les médias et sur Internet, au point que le modérateur du Forum spécialisé LITOR a dû suspendre un débat que la suspicion, la violence et la mauvaise foi avaient dénaturé. Si pour la première fois nous confions à la publication, plutôt qu'au silence, notre idée sur cette affaire, longtemps après avoir été mis en cause, c'est pour garder et défendre la mesure, pour empêcher qu'on ne profite de cet échec pour condamner sans appel la lexicométrie, et même, pour défendre Labbé et son œuvre contre ses propres excès.

Sans être un spécialiste du XVII^{e} siècle, il se trouve que j'ai été amené à m'intéresser, bien avant que Labbé ne s'en préoccupe, des rapports entre Molière et Corneille. Un de mes collègues à l'Université s'était laissé convaincre par la thèse de Pierre Louÿs, en y ajoutant un argumentaire de son propre cru. Un autre collègue de la même université, spécialiste incontesté de la comédie au XVII^{e} siècle, opposait son scepticisme à cette thèse, et l'ordinateur était sollicité de part et d'autre pour une expertise objective. Je fus donc conduit à consulter et à traiter les données du théâtre classique, qui étaient disponibles depuis vingt ans au Trésor de la Langue Française et que Labbé allait reprendre quelques années plus tard, en les complétant. Je m'en suis tenu en effet aux pièces classiques les plus célèbres dont 13 de Molière, 8 de Corneille et 10 de Racine. Or, les trois auteurs dramatiques soumis à un calcul de distance lexicale (le calcul de Jaccard) et

à l'analyse factorielle se détachaient fort bien les uns des autres. Le commentaire de cette expérience se trouve encore aux pages 102-103 du manuel de notre logiciel Hyperbase : « La spécificité des trois écrivains y est excellemment soulignée puisque chacun occupe un coin du graphique. Mais la loi suprême du genre est respectée : le *Menteur* et les *Plaideurs*, tout en s'écartant le moins possible de leur auteur, passent dans le camp de la comédie. »

Cette expérience, déjà ancienne, semblait confirmer les leçons d'une recherche, plus ancienne encore, réalisée avec Charles Muller. Ce spécialiste de Corneille – qui n'a jamais ajouté foi à la thèse de Pierre Louÿs – m'avait proposé un exercice de laboratoire en isolant trois écrivains de la même période et de la même école romantique. En fournissant à l'ordinateur une liste de soixante éléments choisis parmi les mots grammaticaux (on pensait écarter ainsi les aléas thématiques pour mieux cerner les faits stylistiques), nous voulions savoir si les mesures lexicométriques permettraient de reconnaître la griffe de Hugo, de Lamartine et de Musset dans les textes poétiques, romanesques ou dramaturgiques où les relevés avaient été faits. La machine eut beau jeu de reconnaître trois écrivains : un poète, qui avait écrit les *Méditations*, les *Contemplations* et les *Nuits*, un dramaturge qui avait écrit *Lucrèce Borgia* et *Il ne faut jurer de rien* et un prosateur qui était l'auteur de *Raphaël*, de *Notre-Dame de Paris* et des *Confessions d'un enfant du siècle*. Le genre avait malencontreusement recouvert les vraies signatures.

2. UNE EXPÉRIENCE DE LABORATOIRE

Mais les machines et même les hommes ont fait des progrès, et la conclusion négative et presque désabusée des tentatives précédentes n'est peut-être plus de saison. Des outils et des traitements nouveaux sont maintenant disponibles, en particulier ceux que propose Dominique Labbé. D'où l'idée d'une collaboration avec ce chercheur.

2.1. Cependant, pour éviter à la machine une autre humiliation, j'ai cette fois neutralisé le genre. Les textes que la nouvelle expérience met en jeu relèvent tous du genre narratif. En revanche, la variable chronologique, ignorée précédemment, entre en ligne de compte, puisque deux siècles s'interposent entre le texte le plus ancien (*La Vie de Marianne*, Marivaux, 1731) et le texte le plus récent (*Le Temps retrouvé*, Proust, 1927). L'objectif proposé au programme étant de reconnaître la paternité des textes, il suffit, pour chaque auteur, de traiter deux textes qui lui appartiennent et de vérifier si l'algorithme les attribue à la même plume. Pour corser la difficulté, on a choisi pour chaque écrivain d'associer deux œuvres situées aux deux extrémités de sa carrière, pourvu qu'elles partagent le même genre narratif. Il y a ainsi dix-huit ans entre le premier grand succès de Balzac (*Les Chouans*,

1829) et le dernier roman publié de son vivant *(Le Cousin Pons*, 1847). Un laps de temps plus grand encore sépare le premier roman naturaliste de Zola (*Thérèse Raquin*, 1867) et l'un des derniers titres des *Rougon-Macquart* (*La Bête humaine*, 1890). Entre l'un des tout premiers titres de Jules Verne (*De la Terre à la lune*, 1865) et le dernier manuscrit qu'il ait remis à son éditeur Hertzel, quelques jours avant sa mort (*Le secret de Wilhelm Storitz*, 1905), c'est une carrière de quarante ans qui s'est déroulée, modifiant l'inspiration et l'écriture. Cet écart systématique recherché entre les deux spécimens des onze écrivains retenus tendait à dilater au maximum, dans les limites du genre, les différences internes, afin de voir si elles résisteraient aux oppositions externes qui s'exercent entre les écrivains et empêcheraient l'attribution correcte des textes. En somme nous voulions comparer les distances *intra* (entre les textes d'un même écrivain) et les distances *inter* (entre les écrivains).

En réalité, le *nous* collectif que je viens d'utiliser est un abus de langage. Car, j'ai été le seul responsable des conditions de l'expérience et du choix des textes. Dominique Labbé voulait en effet participer à l'expérience en ignorant tout des données, afin qu'aucun préjugé subjectif ne puisse pervertir le traitement. Dans beaucoup de disciplines, l'ignorance est ainsi la garantie de la connaissance, et notre modèle a été le protocole en aveugle que la recherche médicale applique au traitement des malades et au test des médicaments. Or, il y a plusieurs distances possibles, selon qu'il s'agit de deux écrivains différents, ou de deux textes du même écrivain, ou de deux extraits du même texte. On a donc dédoublé tous les textes retenus, afin qu'il y ait pour chaque texte deux extraits différents, mais aussi proches que possible, puisqu'on les a choisis contigus, l'un suivant l'autre. Il y a ainsi pour chacun des onze écrivains quatre extraits qui lui sont attribués, soit 44 au total. Naturellement Dominique Labbé n'a eu droit qu'à des numéros anonymes (pour Proust, c'était 21 et 43, 22 et 44 respectivement). Il n'a pas cherché à les identifier, même si des indices assez clairs − surtout les noms propres − pouvaient aider au décryptage[1]. Pour déjouer toute tentative de cet ordre, un piège avait été tendu dans les six derniers extraits (numérotés de 45 à 50). Car, ce ne sont pas des textes suivis, mais des agrégats constitués de pages empruntées aux 44 textes du corpus, à raison d'une page par texte. Le texte 45 réunit la première page de chaque texte, le texte 46 la dixième, etc. Cela donne des clones qui ne se distinguent pas les uns des autres, mais aussi des portraits robots qui font la synthèse de tous les textes du corpus et en constituent une sorte de moyenne ou d'échantillonnage raisonné. Ce piège a fortement intrigué Labbé, sans l'égarer, et nous invitons le lecteur à lire son commentaire, qui ne manque pas de perspicacité[2].

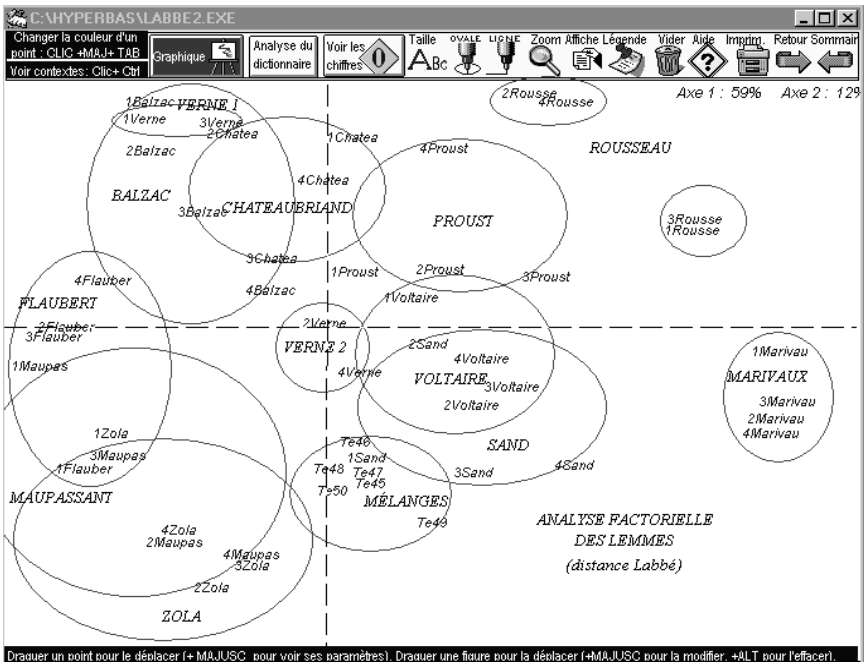


Figure 1 : Analyse factorielle de la distance lexicale
(Formule de Labbé, appliquée aux lemmes)

2.2. Avec les mêmes données et des méthodes semblables aux siennes, nous obtenons les mêmes résultats. Dans la panoplie des outils d'analyse multidimensionnelle, à côté de la classification automatique et de l'analyse arborée, dont D. Labbé a fait usage, on dispose de l'analyse factorielle, qui est illustrée dans la figure 1. Confirmation est donnée du lien très fort qui unit les couples : tout extrait portant l'indice 1 se trouve à proximité immédiate de l'extrait correspondant qui en est la suite et qui est numéroté 3, et il en est ainsi des extraits pourvus des indices 2 et 4. Mais encore les deux couples qui se rattachent au même écrivain ne sont jamais très éloignés, en sorte qu'il est facile de circonscrire dans un cercle plus ou moins étroit les quatre extraits qui relèvent de la même plume. Les concentrations les plus fortes sont le fait des extraits que D. Labbé a désignés comme étant sûrement de la même source : à l'extrême droite les textes de Marivaux (codés 1, 23, 2 et 24) et au centre ce que nous appelons les « mélanges » et qui concerne les extraits de 45 à 50.

Il n'en reste pas moins que l'analyse factorielle, c'est l'aire du soupçon. Elle fournit des présomptions sur une échelle continue qui ne rejoint la certitude que de façon asymptotique. Les certitudes sont parfois positives

(par exemple, le doute n'est guère permis pour Marivaux), mais plus souvent négatives : il est très peu probable que des points diamétralement opposés sur le graphique soient de la même source. Entre ces deux extrêmes on trouve des situations relativement claires et d'autres plus troubles. Parmi les premières on citera les configurations qui tournent autour de Rousseau, Voltaire, Chateaubriand, Balzac et Proust. Mais le troupeau des textes réalistes et naturalistes, à gauche et en bas de la figure, est plus indistinct, comme si les bergers avaient mêlé leurs bêtes. Si Flaubert se distingue assez nettement de Zola, Maupassant évolue librement de l'un à l'autre, plus proche de Flaubert dans *Une Vie*, et de Zola dans *Pierre et Jean*. Le désaccord le plus criant est relatif à Jules Verne : si l'excentricité du premier texte (*De la terre à la lune*, extraits 19 et 41), soulignée par D. Labbé, est bien confirmée par la position extrême (en haut à gauche) des points *1Verne* et *3Verne*, la liaison est rompue avec l'autre texte de Verne qui se situe au centre du graphique (points *2Verne* et *4Verne* recouvrant le *Secret de Wilhelm Storitz*). Ces deux textes de Verne se trouvaient aussi très distants dans l'analyse de D. Labbé. L'explication tient non seulement à la distance chronologique qui sépare les deux textes (40 ans), mais aussi à l'évolution d'un écrivain qui commence par écrire des romans d'aventure pour enfants et qui finit candidat à l'Académie française avec des récits fantastiques et psychologiques écrits à la manière du *Horla* de Maupassant.

En présence du graphique 1, l'œil peut être sensible en outre au mouvement d'ensemble qui, de la droite à la gauche, semble soumettre les textes et les auteurs à la dérive du temps. On observe une sorte de croissant, caractéristique des données sérielles, où prennent place, successivement et dans l'ordre chronologique, Marivaux, Rousseau, Chateaubriand, Balzac, Flaubert, Maupassant et Zola. Cette décantation du temps est pareillement observable dans les deux graphiques de D. Labbé. Rien de très surprenant : en deux siècles la langue a évolué, le mouvement des idées et des sensibilités s'est précipité, et le progrès technique a changé le monde. Pourtant, le courant n'emporte pas tous les écrivains à vitesse constante : il peut se rencontrer des obstacles, des résistances et des remous et certains écrivains semblent remonter le courant. C'est le cas de G. Sand et, plus nettement encore, de Proust qui sur le graphique s'éloigne autant que possible du naturalisme et préfère en haut et à droite la compagnie de Rousseau et Chateaubriand. S'agit-il des thèmes proustiens ou de la phrase proustienne ? Les effets sont mêlés car la mesure proposée par D. Labbé tient compte de la fréquence de tous les vocables, et est sensible aux faits stylistiques, autant que thématiques.

2.3. Pour y voir plus clair et distinguer le thème de la syntaxe, nous avons entrepris d'autres investigations, en poursuivant l'enquête en deçà ou au-delà du lemme. On gardait la mesure de la distance telle que la propose D. Labbé,

mais en l'appliquant à d'autres objets isolés dans le même corpus : des graphies, des codes grammaticaux, des structures syntaxiques ou des étiquettes sémantiques. Mais, d'autres mesures de la distance étaient aussi proposées et comparées à celle de Labbé. Comme on a rendu compte de cette expérience dans une autre publication[3], nous nous bornerons à reproduire la carte des distances établie sur les graphies. Cette fois, nous utiliserons le programme d'analyse arborée que nous avons incorporé à notre logiciel HYPERBASE, parallèlement à l'analyse factorielle de correspondance. La méthode arborée, en effet, est particulièrement adéquate lorsque le tableau à analyser est une matrice carrée, où lignes et colonnes désignent les mêmes objets, et où sont identiques les valeurs lues symétriquement de chaque côté de la diagonale principale (la distance de A à B est la même que de B à A). Les données de la figure 2 sont relatives aux simples graphies, avant toute lemmatisation.

L'interprétation de tels graphes est aisée dans son principe. La distance d'un texte à un autre est directement proportionnelle à la longueur des segments qu'il faut parcourir pour relier les deux points. L'angle, la direction, les tournants et les carrefours n'importent pas, seule compte la longueur du parcours dans un relief tourmenté où les routes empruntent les vallées et les cols.

En partant du haut du graphique, on rencontre d'abord Marivaux dont les quatre extraits sont serrés les uns contre les autres, puis le chemin conduit à Rousseau (mais les deux textes de Rousseau, s'ils débouchent sur la même voie, sont assez distants l'un de l'autre, car il y a loin entre le récit des amours romantiques et l'essai sur l'éducation des enfants). Ensuite, la rencontre de Proust serait inattendue, si nous ne l'avions déjà croisé à cet endroit dans l'analyse des lemmes. Puis, la route hésite ; des voyageurs en retard (Sand, deuxième Verne) ou en avance (Voltaire), ou bien des collectivités indifférenciées (les « mélanges » 45 à 50), ou bien encore un isolé que la naissance a placé au croisement des deux siècles (Chateaubriand), encombrent le carrefour qui conduit à la vallée opposée. Balzac attend là[4], qui passe le relais à Flaubert, puis à Maupassant et enfin à Zola. C'est à peu de choses près le chemin qu'a emprunté D. Labbé, les yeux bandés, en suivant les lemmes.

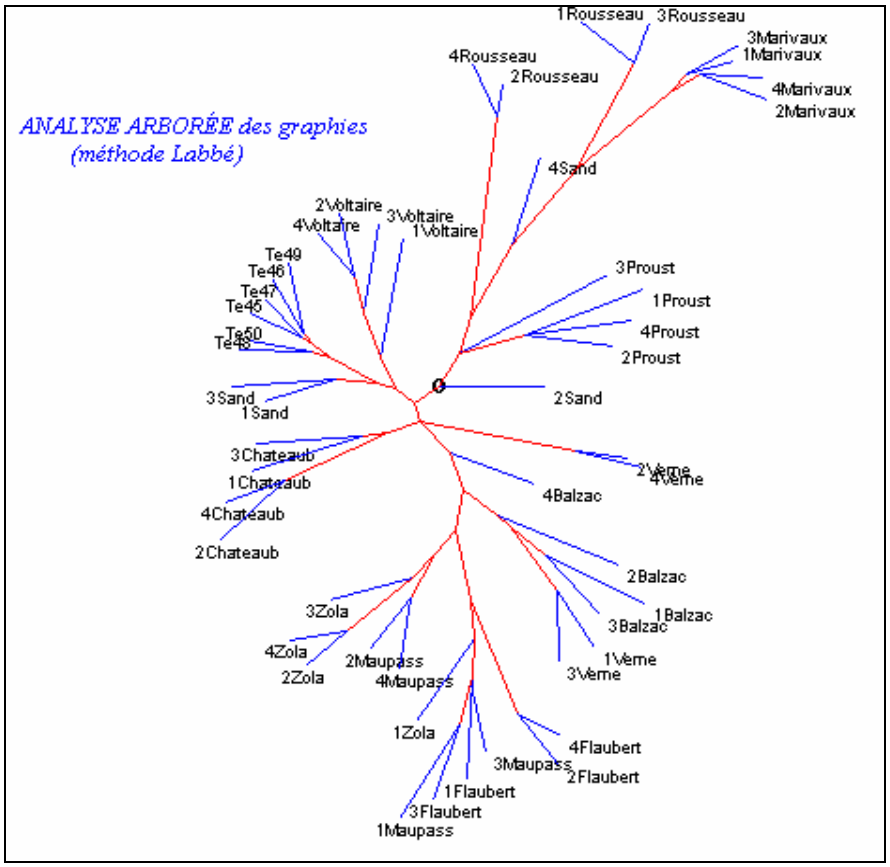


Figure 2 : Analyse arborée (Distance établie sur les graphies)

2.4. Reste à écarter un dernier doute, l'expérience ayant été menée à travers des textes tronqués. De plus, même si le corpus a une taille suffisante, à cause de la multiplication des textes (10 000 x 50 = 500 000 occurrences), chacun des textes traités reste relativement étroit. Les conclusions ne seraient-elles pas plus claires et plus sûres avec des textes complets et une étendue élargie ? La figure 3 répond à cette question en proposant un corpus quatre fois plus vaste (2 millions d'occurrences), constitué des mêmes textes, cette fois sans extraction ni troncature. Reprenons le problème initial et la méthode de Labbé et voyons si la distance lexicale, établie sur les lemmes, pourrait apparier les textes deux à deux et reconnaître une signature commune. Rappelons que les deux textes d'un même auteur ont été choisis à des moments fort différents de la carrière et que rien ne garantit que les thèmes et l'écriture y soient constants. Ils sont pourtant plus proches l'un de l'autre que de tout autre texte. Et cet air de famille est reconnu par l'analyse

arborée qui distribue les couples tout au long de la chaîne. Si les liens familiaux sont prépondérants, l'appartenance à la même époque crée des liens secondaires, de sorte que la procession des couples se fait grossièrement par rang d'âge. Mais, l'ordre chronologique est bousculé à certains endroits, Voltaire se rapprochant de l'époque moderne, tandis que Proust, rompant avec le naturalisme, semble appartenir au siècle précédent. Ces remous dans le fleuve chronologique montrent que le tempérament propre d'un écrivain peut résister au courant et que les procédures d'attribution que nous venons de mettre en œuvre sont plus efficaces contre le temps qu'elles ne le sont contre le genre.

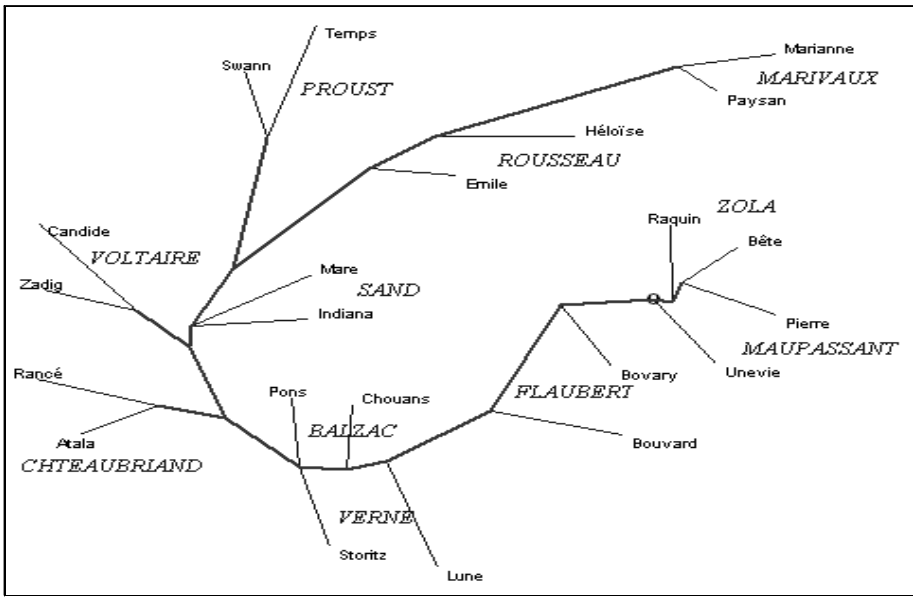


Figure 3 : Analyse de la distance lexicale dans les textes complets
(Les distances sont établies sur les lemmes, selon la méthode Labbé)

3. LIMITES DE LA FORMULE DE LABBÉ

Fort de cette expérience, D. Labbé a cru que la clé pouvait ouvrir d'autres portes et résoudre des problèmes d'attribution plus difficiles que l'exercice d'école qui précède. L'intention est louable car on ne peut réduire toujours la statistique linguistique à un rôle subalterne et ne solliciter son témoignage que lorsque l'affaire est déjà jugée. L'affaire Corneille-Molière – tardivement suscitée par Pierre Louÿs, trois siècles après la mort des intéressés – n'avait pas fait long feu et semblait classée depuis longtemps par les historiens de la littérature. Mais Labbé a fait appel de ce jugement, en invoquant non pas des faits nouveaux mais une méthode d'expertise nouvelle, celle que nous venons de mettre à l'épreuve. Depuis que le recours

à l'ADN est autorisé devant les tribunaux, bien des affaires ont été éclaircies que les témoignages et les autres indices n'auraient pu élucider. ADN, empreintes digitales, carbone 14, ces techniques de dépistage scientifique ont été évoquées dans le procès littéraire où Labbé s'est engagé. Mais la mesure de distance qu'il propose peut-elle jouer ce rôle ? C'est ce que nous nous proposons d'examiner.

3.1. Observons tout d'abord, d'un point de vue théorique, que la statistique peut emprunter deux voies : l'une est inférentielle, l'autre descriptive. La première s'appuie sur les lois probabilistes et permet, à partir d'observations réalisées sur un échantillon, de confirmer ou d'infirmer des hypothèses et de projeter des conclusions sur la population dont l'échantillon est extrait, tout en mesurant la précision et la sûreté de cette projection. La seconde est plus modeste, comme le note le mathématicien Barthélémy, auquel on doit l'analyse arborée et qui s'indigne de l'usage qui en est fait : « Cette utilisation des méthodes que j'ai contribué à mettre au point est un non-sens. On ne peut faire passer pour des statistiques inférentielles, avec lesquelles on peut éprouver des hypothèses, des statistiques descriptives, d'abord destinées à faire réfléchir des spécialistes[5] ». Or, les techniques multidimensionnelles dont on fait usage en lexicométrie, qu'il s'agisse d'analyse factorielle, d'analyse arborée ou de classification hiérarchique, ne sont que des représentations analogiques, qui peuvent fournir des indices, des présomptions, mais non des preuves. Tout est affaire d'interprétation et la nôtre, avec les mêmes données et les mêmes résultats, est assez différente de celle de Labbé, ce que nous montrerons plus loin. L'essentiel du débat – dans la presse comme dans la discussion technique engagée sur la liste LITOR – a porté sur cette prétention de prouver, jugée imprudente et abusive.

Labbé n'a certes jamais caché que sa démarche est empirique, comme celle de tous les chercheurs qui s'adonnent à la lexicométrie, et l'empirisme ne se justifie que par la qualité, l'ampleur et la représentativité des observations. Quoique l'expérience de Labbé soit très large et solide, elle s'est surtout exercée jusqu'ici sur des textes modernes, en relation avec la politique, la sociologie et l'économie. Dans les discours ou entretiens qu'il a étudiés, même parfois sous la plume de de Gaulle ou de Mitterrand, la notion d'auteur a des aspects flous, parce qu'une équipe a souvent préparé ou même rédigé partiellement le texte. Dans le domaine littéraire, la paternité est plus chatouilleuse. On y est sensible aux sources, aux emprunts, aux plagiats, aux querelles d'école, aux contraintes du genre et aux propriétés de l'écriture. Or, Labbé a jusqu'ici rarement exploré ce domaine particulier, sinon dans l'expérience que nous venons de relater. Est-ce assez pour affirmer la valeur universelle d'un test ? Si l'étalonnage de ce test est réellement fondé, comme

on nous l'affirme, sur des « milliers de textes », encore faut-il que la représentativité de ces textes soit assurée. Combien de ces textes appartiennent à la littérature, combien au théâtre, combien à la comédie, combien à la tragédie, combien au genre versifié, combien au XVII[e] siècle ? Labbé se déclare prêt à mener des enquêtes dans ces directions. Que n'a-t-il commencé par là, avant de proposer imprudemment une échelle absolue.

3.2. Nous ne contestons pas l'intérêt de la mesure de Labbé, sans quoi nous ne nous serions pas prêté à l'expérience précédente. Mais, faute d'essais suffisants, nous refusons l'idée d'une échelle fixe, d'un barème arbitraire, attaché à une seule mesure, globale et indifférenciée, appliquée, qui plus est, à un seul aspect – lexical – du langage. Nous croyons même que la formule de Labbé vaut mieux que l'usage qu'il en fait, et nous nous sommes attachés dans les pages qui précèdent à diversifier son emploi, en l'appliquant à d'autres objets linguistiques que le lemme : aux graphies, aux codes grammaticaux, aux structures syntaxiques ou aux réseaux sémantiques. Naturellement, l'échelle des valeurs obtenues varie selon l'objet étudié et le barème pour les lemmes ne vaudrait plus pour les graphies (J. M. Viprey a fort bien observé un décalage approximatif de 4 points[6]). Bien entendu pour les codes et les structures – on pourrait songer aussi aux mesures rythmiques ou prosodiques – l'échelle exigerait des accommodements plus importants. Mais, même dans les conditions précises où se place Labbé, une échelle absolue est impraticable. Elle dépend en effet de certaines options – toutes pareillement justifiables – qui commandent le toilettage du texte, le comptage des mots, et la lemmatisation. Labbé a des exigences particulières quant à la présentation des textes (les hors-texte et didascalies par exemple sont écartés), quant au traitement des mots composés (il en relève un minimum dans les textes classiques), quant à la prise en compte des ponctuations dans le dénombrement des occurrences et surtout quant aux principes de lemmatisation. Le logiciel dont il est l'auteur – et qui a quelque mérite, ayant été construit par un homme seul – se contente d'un codage minimum, qui n'envisage pas la fonction des mots et n'approfondit guère leur nature (ni le temps, ni le mode, ni la personne des verbes ne sont repérés). Il permet cependant de réduire les homographies, à condition que des retouches manuelles viennent suppléer aux embarras de la machine. Ces retouches évitent certes bien des erreurs grossières qu'on constate dans les résultats des lemmatiseurs automatiques, comme ceux de Cordial. Mais le prix à payer en temps est élevé, sans garantir la constance des décisions, qui varient d'un chercheur à l'autre, et parfois même d'un moment à l'autre. Sauf à confier à Labbé le traitement de tout texte que l'on veut soumettre à son calcul de distance (d'autant que son lemmatiseur n'est pas commercialisé), on voit mal comment on pourrait appliquer son échelle, si

les conditions de mesure ne sont pas semblables. Tous les linguistes appellent de leurs vœux une standardisation minimale dans la saisie, le codage, la lemmatisation et le traitement des textes, mais cela ne peut résulter que d'un consensus international fixant des normes précises (ce que l'entreprise de *Text Encodage Initiative* s'emploie à réaliser, d'autant que le codage *XML* en donne les moyens), ou à tout le moins sur une tradition nationale – qui en France est représentée majoritairement par *FRANTEXT* et l'*Institut de Linguistique Française*. Toute tentative individuelle, même excellente, est vouée à l'échec.

La conséquence de cette situation est que les méthodes et les résultats de Labbé sont infalsifiables, puisqu'on doit passer par lui pour les approuver ou les combattre. Il est certes facile de trouver des contre-exemples où le barème invite à considérer deux textes comme appartenant à la même plume, alors qu'on sait de façon sûre qu'il n'en est rien. Mais, Labbé peut toujours les récuser, en prétendant que les conditions du calcul n'ont pas été remplies, puisqu'il est le seul à pouvoir les remplir. Dans une base publiée il y a cinq ans, et distribuée par l'Éducation nationale sous le nom de *Batelier*, nous avions appliqué le calcul de Labbé à une soixantaine de textes, dont le *Menteur* et une trentaine de pièces classiques. Certes, la proximité du *Menteur* (et aussi de l'*Illusion Comique*) avec les pièces en vers de Molière y avait été observée, mais aussi celle des *Fleurs du mal* et des *Poésies* de Rimbaud. À l'époque, la formule de Labbé n'avait pas les correctifs qu'elle a reçus depuis et nous lui en avions ajouté un (en refusant les hapax non seulement du texte le plus long, ce que recommande Labbé, mais aussi du plus court). Et bien entendu nous ne disposions pas de la lemmatisation Labbé. Le résultat (0,182) n'a donc pas à être confronté à l'échelle établie depuis lors mais aux autres résultats obtenus dans le même corpus, avec les mêmes options et les mêmes conditions. Or cette proximité entre les recueils de Baudelaire et de Rimbaud est aussi étroite que celle qui lie au *Menteur Don Juan* (0,180), le *Misanthrope* (0,173), l'*Avare* (0,177), les *Femmes savantes* (0,173), le *Bourgeois gentilhomme* (0,222) et le *Malade imaginaire* (0,207). Si donc on conclut que l'auteur du *Menteur* est le même que celui des pièces citées, on doit pareillement conclure qu'il n'y a qu'un auteur pour les *Poésies* rimbaldiennes et les *Fleurs du mal*[7].Ce contre-exemple n'est d'ailleurs pas le seul que nous ayons relevé : en réunissant dans une même base l'œuvre de Molière et celle de Marivaux, les calculs de distance montrent bien une séparation nette entre les deux dramaturges, à l'exception de la première pièce de Marivaux, qui, il est vrai, est fort courte et la seule qu'il ait écrite en vers. Sans doute aussi s'inspire-t-elle du grand devancier mais elle n'est pas de Molière, malgré les indications du barème. Nous avons pareillement réuni l'œuvre de Flaubert et celle de Maupassant, et là encore le seuil de fusion est atteint pour *Madame Bovary* et *Une vie*. Il l'est aussi si

l'on compare les quatre évangiles dans trois traductions françaises qui en ont été faites, soit douze versions différentes. Le calcul semble indiquer un auteur unique, qu'il s'agisse ou non du Saint Esprit.

3.3. Une autre raison invite à renoncer au barème proposé par Labbé, c'est l'obscurité qui s'attache à une mesure unique et globale. Ce que l'on gagne en synthèse est perdu en analyse. Comment en effet interpréter une mesure de proximité quand plusieurs facteurs sont en cause. Labbé est sensible à cette difficulté et il détaille les influences qui entrent en ligne de compte : l'auteur, le genre, le sujet, l'époque. Mais dans une mesure donnée, rien ne permet de distinguer ces influences variables, dont le dosage échappe au calcul. Dès lors ce qu'un chercheur interprète comme caractéristique d'un écrivain, un autre critique peut l'attribuer aux contraintes exercées par le genre, voire aux lieux communs que le sujet entraîne. Les cas où le calcul est opérant sont ceux où les variables indésirables sont neutralisées. Dans une émission sur France-Culture, Labbé opposait à ses contradicteurs le cas de *Tite et Bérénice* où son calcul fait merveille pour distinguer la pièce de Racine et celle de Corneille. On aurait pu lui répondre que les conditions idéales étaient réunies (même sujet, même année et même genre) pour rendre le calcul efficace et explicite, mais qu'elles ne l'étaient plus dans le cas Corneille-Molière qui faisait l'objet du débat. Quoi de plus attendu que la proximité du *Menteur* et des pièces de Molière ? Ce sont des comédies et celles qui sont les plus proches sont celles qui, comme le *Menteur*, sont écrites en vers. La seule comédie que Racine ait écrite, les *Plaideurs*, est également plus proche de Molière que de Racine. Pourquoi ne pas se contenter de ces remarques de bon sens ? Pourquoi s'ingénier à chercher une explication hypothétique du côté de l'auteur, en refusant le facteur le plus évident, c'est-à-dire le genre[8] (d'autant que le genre est très contraignant à l'époque classique où de surcroît la versification impose des exigences supplémentaires) ? Devant l'impossibilité de démêler des facteurs entrecroisés et indissociables, le principe de précaution est de ne pas parler de preuve et de laisser à Pierre Louÿs le soin de défendre sa rêverie et ses intuitions.

3.4. Reste à apprécier en elle-même la formule par laquelle Labbé mesure la proximité entre deux textes. Nous préférons le terme de proximité à celui de distance. Car la distance, notion familière et abstraite dans l'esprit des mathématiciens, peut prêter, dans d'autres esprits, à des confusions engendrées par la métaphore géographique. Et il arrive à Labbé de tomber dans ce piège : « La distance est une mesure physique. Par exemple, St-Germain-en-Laye […] et Paris forment aujourd'hui une seule agglomération alors que Rouen est suffisamment éloignée pour être considérée comme une

entité urbaine distincte […]. Il est absurde d'objecter à cela que nous devons d'abord mesurer Paris-Lyon, Paris-Lille… et pourquoi pas : Paris-Oulan Bator[9] ? » Or, la distance entre deux textes, c'est comme la proximité entre deux êtres ou deux cultures : elle suppose d'autres textes, plus ou moins proches, un espace où les accointances ou répulsions réciproques puissent se déployer. La distance intertextuelle est relative et n'a pas de sens si les points de repères manquent.

Et surtout, la distance est multiple. Il y a bien des façons de rapprocher deux textes ou deux objets. Les mathématiciens en ont inventé des centaines. Et, il y a une chance que celle qu'on croit trouver a déjà été imaginée par quelqu'un d'autre. Ainsi, nous avons eu la surprise de retrouver récemment dans une revue datant de 1989[10], la formule de Jaccard que nous avions aménagée à notre façon pour la rendre indépendante de l'étendue. Cette formule figure avec vingt autres, pareillement justifiées, et toutes établies, non sur la fréquence, mais sur la présence/absence. Nous en avons profité pour modifier notre calcul en empruntant à cette source un quatrième ingrédient jusqu'ici négligé : le nombre de mots qui ne figurent dans aucun des deux textes comparés. Car, la proximité peut résulter non seulement de goûts communs, mais aussi de dégoûts partagés. En taxinomie, s'il est utile de connaître les propriétés qui appartiennent aux deux éléments comparés, et celles qu'on ne trouve que dans un seul, il n'est pas sans intérêt de savoir celles qui sont exclues de part et d'autre. En lexicométrie, le calcul exige évidemment qu'on soit enfermé dans un corpus fini, afin qu'on puisse dénombrer les mots qui manquent dans la confrontation de deux textes mais qu'on rencontre dans les autres.

3.5. Les calculs de proximité qui font intervenir la fréquence sont beaucoup moins nombreux. Et c'est pourquoi nous avons porté un intérêt à la formule de Labbé. On dispose certes de deux procédures dont l'une remonte à Muller. Dès 1968, dans son *Initiation à la statistique linguistique*[11], Charles Muller proposait l'application de la loi binomiale au calcul de ce qu'il appelait la connexion lexicale. Ce calcul reposait sur les classes de fréquence, et donc éliminait complètement la composante sémantique et thématique des textes. Nous renvoyons le lecteur aux deux applications que nous en avons faites, à propos de Giraudoux et de Hugo, et qui sont, à notre connaissance, sans autre exemple. La chaîne des calculs y est en effet fort longue et, si elle aboutit à un Chi2 synthétique qui évalue la proximité des deux textes comparés, elle nécessite une pondération qui amortisse l'effet des grands nombres, et donc de l'étendue des textes, sur toute mesure probabiliste[12]. La seconde méthode est en revanche très connue, très classique et très rapide. Elle est recommandée par A. Salem et J. M. Viprey, au moins pour une première approche. Il s'agit tout bonnement de l'analyse factorielle appliquée au TLE (tableau lexical entier), c'est-à-dire au

dictionnaire des fréquences et sous-fréquences, que les logiciels d'indexation construisent tous à un moment ou à l'autre du traitement. En réalité, le TLE est rarement proposé en entier, car les calculs, peu légitimes dans les basses fréquences, allongeraient exagérément le nombre de lignes du tableau. Mais, l'algorithme étant très rapide, des tableaux de quelques milliers de lignes (c'est-à-dire de mots différents) sont traités en quelques secondes.

3.6. La formule de Labbé apporte un heureux complément à la méthode précédente. Elle est plus sensible aux fréquences basses qu'aux mots fréquents, les premières accaparant 40% de la distance totale quand les seconds, pour une surface avoisinante, ne rendent compte que de 5% de la variance. On trouvera sur ce point, dans la revue *Corpus* (n°2, *La distance intertextuelle*, Nice, décembre 2003), la mesure détaillée que Labbé fait de la contribution des différentes classes de fréquence (et aussi des parties du discours). L'explication qui en est donnée ne nous convainc qu'à moitié : les hautes fréquences seraient plus régulièrement distribuées que les basses, mis à part quelques mots très sensibles à la situation du discours comme les pronoms personnels. En réalité l'influence prépondérante des basses fréquences vient de leur nombre. Comme il y a un vote par mot, rare ou fréquent, pauvre ou riche, la voix des puissants se perd dans la rumeur du peuple. La démocratie égalitaire y a pourtant ses limites. Labbé recommande d'éliminer les hapax et plus précisément les mots rares qu'on rencontre dans le texte le plus long et dont la fréquence théorique dans le plus court serait inférieure à 1. Il invite aussi à ne pas tenir compte des écarts inférieurs à 0,5. Ces retouches sont probablement fondées en pratique, mais elles affaiblissent la pureté de la formule et, en limitant la population appelée à voter, elles diminuent un peu le crédit de la consultation. La formule de Jaccard au contraire est dénuée de rustines et d'emplâtres. Tous les mots, hapax compris, sont invités aux urnes, même si le vote de certains est connu d'avance : les mots très fréquents ne peuvent éviter de se trouver dans la zone commune.

On comprend mieux maintenant la convergence, très souvent observée, des mesures de Jaccard et de Labbé. Quoique l'une s'attache à la simple présence et l'autre à la fréquence, toutes les deux rendent compte en priorité des basses fréquences. Et toutes les deux ont à lutter contre les perturbations que l'étendue inégale des textes peut provoquer. Elles y réussissent certes, mais imparfaitement. Considérons en effet les 75 textes du corpus classique réunissant les pièces de Corneille, Molière et Racine. Cela fait 2775 distances à calculer, soit n*(n-1)/2. Un tri sur la longueur des textes (ou plus exactement sur le rapport d'étendue des textes deux à deux) met en relief une légère distorsion, de même nature, dans les deux procédures. Dans les deux mesures, la distance est plus faible, et plus fiable, quand les textes sont de

longueur voisine et que le rapport d'étendue s'approche de 1 (de 100 sur le graphique 4).

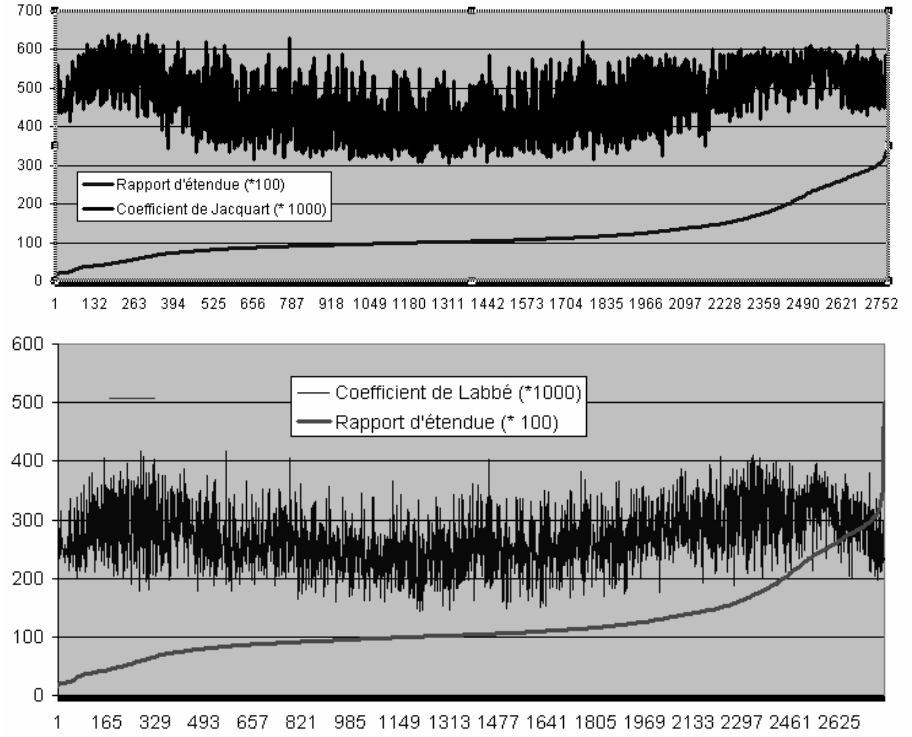


Figure 4 : L'influence de l'étendue sur les distances de Jaccard et de Labbé.

Cette convergence des deux approches – jusque dans les défauts – est rassurante et c'est pourquoi nous avons tenu à mettre en parallèle les deux programmes de distance dans notre logiciel. Le programme de Jaccard est d'une simplicité et d'une rapidité extrême quand on dispose du TLE. L'algorithme de Labbé est pareillement simple à mettre en œuvre, mais le temps du calcul est nettement plus long et augmente exponentiellement avec le nombre de textes du corpus. Aussi, avons-nous rendu cette fonction facultative. Quelques lignes de code suffisent à la traduire : on les déchiffrera aisément, si l'on sait que pour un mot donné les sous-fréquences sont cataloguées dans le tableau TABLE, *nb* étant le nombre de textes du corpus, tandis que *dista* et *disma* reçoivent la sommation du numérateur et du dénominateur de la formule. La boucle est à répéter pour chaque lemme.

step k from 1 to nb-1 step l from k+1 to nb if table[k] = 0 and table[l] = 0 continue step	end if taille [k] < taille [l] coef = taille [k]/taille[l] theo = table [l] * coef	if table [k] = 0 and theo < 1 continue step end ecart = abs (table [k] - theo)
if ecart < 0.5 continue step end else coef = taille [l]/taille[k] theo = table [k] * coef if table [l] = 0 and theo < 1	continue step end ecart = abs (table [l] - theo) if ecart < 0.5 continue stepend end. dista[k][l]= dista[k][l]+ecart	disma[k][l]=disma[k][l]+(table[l]+ theo) end end

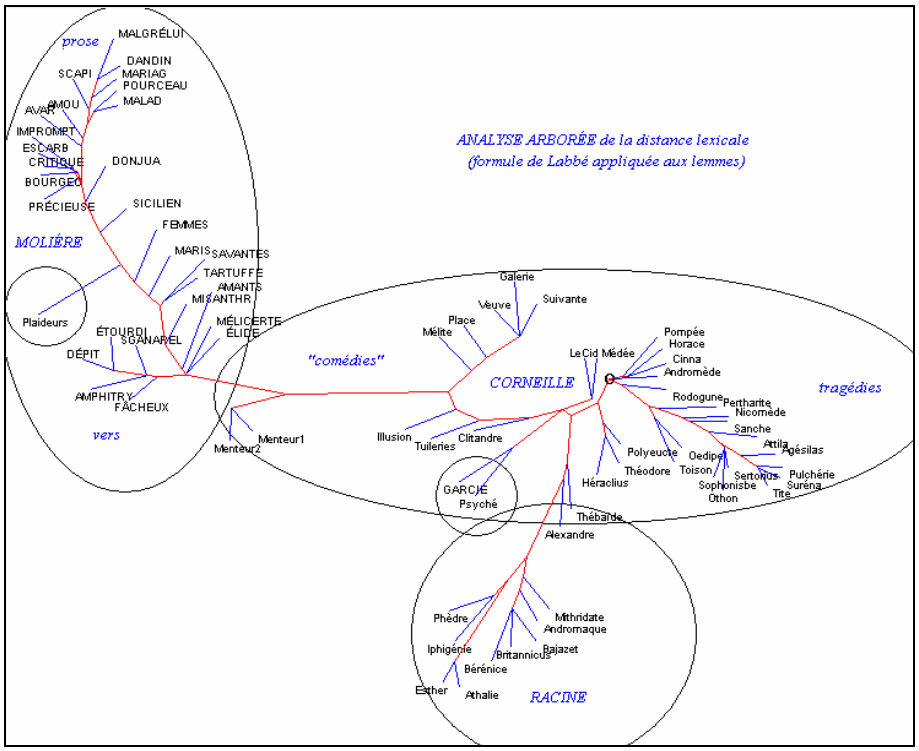


Figure 5 : Analyse arborée de la distance lexicale (méthode Labbé appliquée aux lemmes)

4. LE CAS CORNEILLE-MOLIÈRE. UN PROBLÈME D'INTERPRÉTATION

Après avoir tenté d'évaluer et de comparer les mérites et les limites des mesures de distance, reste à interpréter leurs indications. Et si jusqu'ici nous avons suivi et assez souvent approuvé la démarche de Labbé, tout en refusant son barème, nous nous en séparons radicalement au moment crucial de l'interprétation. Dans le cas du théâtre classique, les résultats pour qui sait les lire sans idée préconçue n'invitent nullement à conclure que Corneille aurait écrit les chefs-d'œuvre de Molière. Bien au contraire, la mesure de Labbé tendrait plutôt à distinguer les deux écrivains. Considérons en effet la carte des proximités, telle que la dessine l'analyse arborée (figure 5).

Il est facile d'en détacher la branche Racine (si on peut dire) qui se dégage mollement d'abord de l'influence de Corneille (la *Thébaïde* et *Alexandre* sont proches du grand devancier) puis affirme son indépendance. On y distingue même la rupture qui, à partir d'*Iphigénie,* conduit l'auteur à *Athalie*. Une telle finesse dans le détail a tout pour plaire aux exégètes les plus exigeants. Mais ne cherchons pas là la comédie des *Plaideurs*. Personne n'a mis en doute son authenticité. Mais comme elle relève d'un autre genre, le calcul a déplacé cette pièce très loin sur la gauche, au beau milieu des comédies de Molière. On chercherait vainement une autre explication : le genre est ici prédominant.

De la même façon, la seule pièce sérieuse qu'ait écrite Molière, *Dom Garcie de Navarre*, a déserté la moitié gauche, où toutes les comédies de Molière sont rassemblées, pour se fixer dans le camp opposé, parmi les tragédies. Est-ce suffisant pour prétendre que Corneille (ou Racine) ait écrit cette pièce ? Le genre suffit à expliquer ce déplacement, comme celui de *Psyché*, qui se situe au même endroit, et dont le genre hybride (tragédie-ballet) est également éloigné de la comédie[13]. Comme Molière et Racine n'ont guère exploité qu'un seul genre, mis à part ces trois exceptions, leur individualité est fort bien circonscrite par le calcul et toutes les comédies de Molière campent à gauche, les pièces en prose en haut et les pièces en vers en bas[14], tandis que les tragédies de Racine sont serrées les unes contre les autres dans le quadrant inférieur droit. Il n'en est pas de même avec Corneille dont l'œuvre est plus diversifiée, plus étalée dans le temps et qui s'est illustré dans plusieurs genres. La surface que le calcul lui attribue est plus large, plus aplatie, et répartie en deux zones : celle des tragédies à droite et celle des pièces comiques ou assimilées à gauche. Il n'en reste pas moins que l'originalité des trois auteurs est préservée, malgré la polarisation du genre. Même les pièces de Corneille forment un bloc, dans lequel entrent les deux *Menteurs*. Les *Menteurs* se rapprochent certes des pièces en vers de Molière, près de la frontière. Mais ce sont des frontaliers, non des transfuges. Quant aux pièces de Molière, aucune ne se compromet avec les pièces de

Corneille. Et l'on comprend mal que Labbé, au vu d'un tel graphique, ait pu les attribuer à Corneille. En réalité au lieu de considérer le jeu d'en haut, d'un regard impartial et neutre, Labbé, barème et baromètre en mains, s'est introduit dans la partie, en privilégiant un ou deux joueurs parmi les 75 en jeu. En focalisant son attention sur les *Menteurs*, qui se situent à la frontière, il a rassemblé sous le même drapeau tous ceux qui se trouvaient dans le voisinage, et les a soumis au même suzerain (il a choisi Corneille, mais Molière aurait pu tout aussi bien revendiquer la conquête en annexant à son territoire les comédies de Corneille, de *Mélite* à l'*Illusion comique)*. L'erreur d'interprétation réside dans ce parti pris que rien ne justifie. Quand on a 2775 mesures de proximité à synthétiser, cela ne peut se faire qu'en prenant du recul, pour les embrasser du regard sans en fixer aucune en particulier. Les méthodes multidimensionnelles (l'analyse factorielle des mêmes données est aussi claire) servent précisément à élargir le champ de la vision en évitant la myopie, et à faire apparaître dans le paysage les massifs et les lignes de partage.

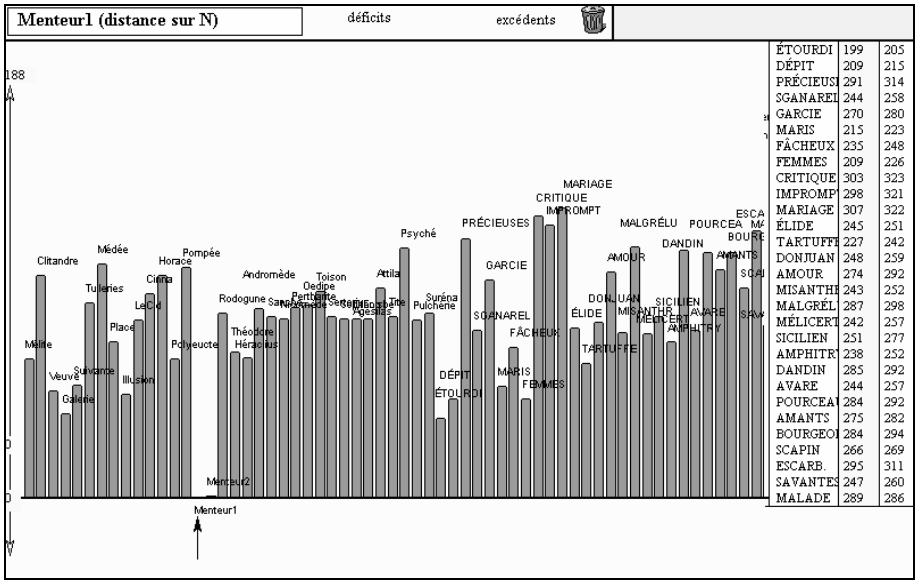


Figure 6 : Distance du Menteur aux 74 autres pièces

Au besoin, avant ou après cette synthèse, rien n'interdit de concentrer son attention sur une ligne ou une colonne du tableau, par exemple celle qui correspond au *Menteur*, comme dans la figure 6. On constate en effet que cette comédie a des accointances fortes non seulement avec les autres comédies de Corneille, mais aussi avec celles de Molière, pourvu qu'elles soient en vers. Et, comme on l'a vu avec les deux premières pièces de Racine, l'influence de Corneille est la plus forte au début de la carrière, dans

les premiers essais de Molière, l'*Étourdi* et le *Dépit amoureux*, ce qui n'en fait pas nécessairement des chefs-d'œuvre[15]. Ce gros plan sur une pièce est certes riche d'informations, mais les 74 autres contiennent autant de renseignements, parfois concordants, parfois divergents. La difficulté des taxinomies et des calculs de proximité vient de l'absence de transitivité. Si A ressemble à B et à C, il ne s'ensuit pas que B ressemble à C. C'est le nœud gordien des 2775 coefficients entrelacés qu'il faut dénouer et il ne suffit pas de tirer sur un fil.

Beaucoup d'autres analyses viennent renforcer l'interprétation qui s'impose dans la figure 6[16]. Celle qui suit (figure 7) reprend le même corpus en lui appliquant un calcul de distance différent, expliqué précédemment sous le nom de Jaccard. Il faut bien se persuader que le programme d'analyse arborée place automatiquement tous les textes, en s'arrangeant pour que s'assemblent ceux qui se ressemblent, comme ferait avec ses invités une maîtresse de maison avisée. Les routes et les chemins sont également tracés, de sorte que le travail d'interprétation ne consiste guère qu'à reconnaître, circonscrire et désigner les agglomérations. Elles sont trois, là encore, et faciles à nommer : la première s'appelle Racine (en haut), la seconde Corneille (au centre) et la troisième Molière (en bas). Impossible de répartir autrement la population. Les trois circonscriptions sont indépendantes et franchement séparées. Si le résultat avait ressemblé à la carte des Balkans, avec des ethnies dispersées et entremêlées, le regroupement aurait pu se justifier. Mais ici, tout est en ordre et les trois écrivains règnent sur des terres que nul ne conteste (mis à part les trois exceptions qu'on a relevées précédemment et qui jouissent de l'exterritorialité du genre littéraire). Ce n'est pas que le genre s'efface. On voit bien qu'il suggère une bipartition : toutes les tragédies sont en haut, et toutes les comédies en bas, et cela sans aucune exception. On voit aussi qu'une décantation se fait qui, chez Corneille, ne mêle pas les comédies et les tragédies et, chez Molière, les vers et la prose. On voit enfin que d'un bout du graphique à l'autre une hiérarchie s'établit entre les pièces : le théâtre d'en bas, c'est la comédie en prose, celui d'en haut, c'est la tragédie en vers, et entre les deux c'est la comédie en vers, que Corneille et Molière se partagent.

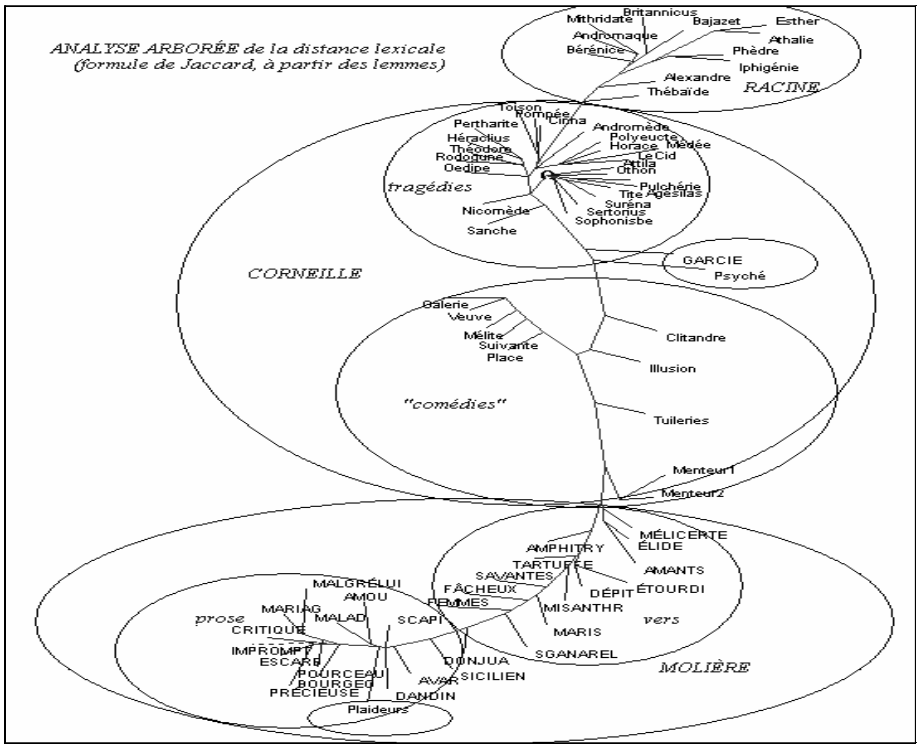


Figure 7 : Analyse de la distance Jaccard (appliquée aux lemmes)

Le plus surprenant est peut-être que l'aimantation du genre, si puissante qu'elle soit, n'ait pas dominé davantage la personnalité des trois écrivains et que le territoire de chacun soit si nettement délimité.

Les historiens de la littérature nous ont appris que leur entente a été médiocre et que chacun avait sa fierté, sa personnalité, ses ambitions, ses jalousies et aurait mal supporté qu'on lui fasse de l'ombre. Et chacun a son originalité très reconnaissable sur le graphique. Ainsi, bien loin de conforter la thèse de Pierre Louÿs, la statistique paraît plutôt l'infirmer[17].

5. NOTES

(1) Il ne serait pas sans intérêt de confronter la sagacité humaine à l'expertise de la machine. La lecture humaine, armée de connaissances externes et attentive aux indications du texte, devrait pouvoir reconnaître les extraits qui vont ensemble. Mais cela dépend de la culture du lecteur. Un ignorant fera des erreurs. On verra dans l'analyse de Dominique Labbé que l'ordinateur, appuyé sur les seuls comptages, et dénué de toute culture, ne se trompera pas une seule fois dans l'identification des couples.

(2) Cet article se trouve sur le site de l'auteur à l'adresse : http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/Labbe.html.

(3) Actes des Troisièmes Journées de la linguistique de corpus, Lorient, 2003, sous presse.

(4) Le premier Verne, celui de la lune, aussi.
(5) Cité dans un article du journal Le Monde, du 10 juin 2003, sous la signature de Fabienne Dumontet, (Molière et Corneille confondus).
(6) Dans le cas du corpus Molière-Corneille-Racine, le décalage moyen est de 0,031, pour 75 textes et 2775 mesures.
(7) En reprenant les mêmes données avec la formule exacte, les conclusions sont les mêmes : le coefficient pour Baudelaire-Rimbaud (0,296) est du même ordre que les autres (respectivement 0,289 0,271 0,289 0,277 0,332 0,323).
(8) Labbé me prête des propos que je n'ai jamais tenus (« le genre est tout, l'auteur n'est rien »). Il s'obstine aussi à m'attribuer, pour aussitôt la contester, une formule qui m'est parfaitement étrangère et qui ne se trouve nullement dans l'article qu'il cite. La formule que je propose depuis dix ans pour mesurer la distance lexicale est dérivée de celle de Jaccard. Elle s'écrit comme suit :
d = ((a-ab)/a) + ((b-ab)/b)
où ab désigne la partie commune aux vocabulaires a et b (a-ab et b-ab recouvrant les parties privatives). Nulle part on n'y fait intervenir la fréquence des mots (Fia et Fib) et la taille des textes Na et Nb, ingrédients de la formule qu'on m'attribue et qui se trouve répétée une fois de plus dans l'article de Labbé « Inter-textual distance and authorship attribution » (Journal of Quantitative Linguistics, 2001, vol 8, n° 3, p 215). Enfin pour en finir avec les allégations inexactes, aucun logiciel digne de ce nom ne traite différemment les minuscules et les majuscules qu'on trouve en tête de vers (ou en tête de phrase). Labbé croit voir là un défaut qui discrédite les travaux lexicométriques portant sur les vers. Tous les utilisateurs d'Hyperbase savent que la majuscule est neutralisée dans l'indexation et les traitements. Ils y trouveront aussi (p. 58 du manuel) la formule de Jaccard que le logiciel exploite et que je viens d'expliciter.
(9) D. Labbé, Réponse à mes contradicteurs, à l'adresse Internet :
http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/Labbe.html
(10) F.B. Baulieu, « A classification of Presence/Absence Based Dissimilarity Coefficients », Journal of Classification 6:233-246 (1989).
(11) Ce manuel, publié d'abord chez Larousse, puis, en deux volumes, chez Hatier, est maintenant au catalogue de Champion.
(12) Le Vocabulaire de Giraudoux. Structure et évolution, Slatkine, 1978, p.369-396. Le Vocabulaire de Victor Hugo, Slatkine, tome 1, p. 277-305.
(13) Dans le cas de Psyché, une raison supplémentaire s'ajoute à l'influence du genre : si la pièce figure bien parmi les œuvres de Molière qui en a créé et développé le canevas en prose, la versification en a été faite, en grande partie, par Pierre Corneille, comme la version versifiée de Don Juan, réalisée après la mort de Molière, est due à Thomas Corneille.
(14) L'influence du genre peut être complexe, car la notion de genre, comme l'a bien montré Rastier, admet des sous-catégories. À un certain niveau le choix se fait entre comédies et tragédies. Au niveau supérieur, on devrait choisir entre théâtre, roman, correspondance, essai, etc… Au niveau inférieur deux options se présentent, vers ou prose, au moins pour la comédie (car il y a peu d'exemples de tragédies en prose au XVIIe siècle).
(15) Les distances, multipliées par 1000, servent d'ordonnées à la représentation graphique. Elles sont lisibles dans les deux colonnes de droite. Celles que Labbé a publiées partiellement sont dans la dernière. On les comparera aux nôtres qui apparaissent dans l'avant-dernière et qui ont été calculées avec le même algorithme mais en tenant compte des ponctuations et des hors-texte et en les soumettant à la lemmatisation de Cordial. Nos chiffres sont légèrement et constamment inférieurs, de 1% en moyenne, ce qui n'a aucune influence sur l'analyse.
(16) Elles portent sur les graphies, les parties du discours, les structures syntaxiques, la segmentation de la phrase, la longueur des mots, les classes de fréquence, etc. La convergence est au rendez-vous mais la place nous manque pour développer ces points de vue. On est loin

d’avoir tout dit sur un texte quand on a fait le relevé des lemmes. Bien d’autres aspects doivent être envisagés, qui font intervenir la syntaxe, la thématique, la métrique.

(17) Poussé par un scrupule de dernière minute, nous avons soumis le corpus à un autre calcul de distance, connu sous le nom de corrélation de Bernouilli et proposé par Étienne Évrard dès 1966 (« Étude des dialectes bantous », in Statistique et analyse linguistique, PUF, p.85-103). Ce coefficient et sa variante simplifiée sont du type Jaccard (les relevés portent sur la présence/absence) et figurent en bon rang dans la liste établie par Baulieu (voir note 12). Les résultats sont tout à fait superposables à ceux du graphique 7.

ANALYSE LEXICOMÉTRIQUE DE L'OPPOSITION GÉNÉRIQUE DANS UNE PERSPECTIVE ENDOGÈNE.

Margareta Kastberg Sjöblom
ILF − CNRS Bases, Corpus et Langage UMR 6039

1. INTRODUCTION

La notion de genre, et notamment de genre littéraire, reste encore aujourd'hui l'institution première du code littéraire, bien qu'elle ait souvent été discutée et mise en question. Les théoriciens la considèrent avec réserve, affirmant que chaque genre littéraire en englobe plusieurs : la nouvelle peut se présenter en effet sous forme de fable, de lettre, de poème en prose, etc. Les hésitations terminologiques (nouvelle, conte, récit…) manifestent ce caractère "d'appartenance multiple et emboîtante" de tout écrit littéraire. En effet, la codification des genres n'est pas chose aisée ni stabilisée. Le système traditionnel nous propose – ou nous impose – selon le code générique institutionnel, certaines classifications reconnues : romans, nouvelles, essais, etc.

Pourtant les études ont montré que les genres existent, qu'on le veuille ou non, et qu'il serait inconcevable sur le plan purement linguistique de nier l'existence des différentes typologies de textes. Par ailleurs, dans l'étude lexicométrique l'opposition générique est extrêmement claire et permet de définir des caractéristiques génériques en s'appuyant, non sur des valeurs culturelles ou sociales, mais sur les propriétés mêmes des textes.

Les études lexicométriques et l'analyse du corpus en situation montrent en effet que le lexique, la morphosyntaxe, la structure et la longueur des phrases, entre autres, varient avec les genres. L'opposition entre les différentes typologies est toujours présente et souvent même prépondérante dans les différentes analyses statistiques. Cette opposition générique dans les analyses de statistique lexicale est si forte qu'elle empêcherait même de fonder de grands espoirs sur les méthodes quantitatives pour attribuer un texte à un écrivain plutôt qu'à un autre.

Ces variations, indiscutables et déjà bien documentées par ailleurs, sont-elles observables aussi à l'intérieur d'un corpus ou dans l'œuvre d'un seul écrivain ? Comment évoluent-elles ? Nous nous proposons ici d'étudier les variations et les oppositions génériques dans une perspective endogène de

l'œuvre d'un de nos plus grands écrivains contemporains : J. M. G. Le Clézio.

Cet auteur s'est lui-même intéressé à tout le procédé de la création littéraire et ses idées se traduisent souvent par un refus de certaines normes littéraires, se présentant comme une contestation sociale. Accepter les conventions du roman, ou de tout autre type d'écriture, présentait pour l'écrivain, surtout au début de sa création, le risque de s'enfermer dans un système sociopolitique, dans un cloisonnement conventionnel des genres qui le dérangeait au plus profond. Tout au long de sa production littéraire, Le Clézio a en effet tenté des expériences en transgressant les catégories et les genres, en écrivant "ni essais, ni romans, ni poèmes, et pourtant tout cela à la fois" (Le Clézio, 1969).

Pourtant, même en voulant transgresser un système social établi, les différentes typologies de textes restent présentes et leurs variations sont à observer à tous niveaux. Les oppositions génériques sont premièrement à constater dans la structure du vocabulaire et dans son évolution ; c'est l'étude de la richesse lexicale, de la diversité du vocabulaire, de l'accroissement lexical ainsi que des hapax qui permet de tirer des conclusions de ce phénomène.

L'étude des parties du discours et de la syntaxe à travers une analyse "grammatico-métrique", possible grâce aux versions lemmatisées et étiquetées du corpus, permet de relever aussi certains aspects morphologiques et syntaxiques qui différencient les types de textes.

L'opposition générique opère aussi au niveau du contenu lexical et thématique d'une œuvre. L'étude de la distance lexicale entre les différents livres du corpus, ainsi que celle des spécificités lexicales, met en exergue les variations thématiques, ou isotopiques, récurrentes dans ce corpus "multigénérique".

2. LE CORPUS

La production littéraire de Le Clézio est vaste, s'étend sur plus de quarante ans et englobe plusieurs genres littéraires. Le corpus informatisé, dans la version qui s'appuie sur les formes graphiques, contient 2.281.659 occurrences et 51.009 formes réparties sur les trente et une œuvres du corpus.

Le corpus est constitué tout d'abord des six premières œuvres, classées par leur style particulier et innovant, comme appartenant à l'École du "nouveau roman" : *Le procès-verbal*, les nouvelles de *La fièvre*, *Le déluge*, *Le livre des fuites*, *La guerre* et *Voyages de l'autre côté*. Les romans qui suivent cette période, considérés par les critiques comme plus "traditionnels", sont au nombre de neuf : *Désert*, *Le chercheur d'or* et

Voyage à Rodrigues écrit sous forme de journal personnel, *Angoli Mala*, *Onitsha*, *Etoile errante*, *La quarantaine*, *Poisson d'or*, et *Hasard*.

Mydriase et *Vers les icebergs* sont difficiles à classer dans un genre précis, ce sont plutôt des récits poétiques. Lorsque certaines critiques les rapprochent de la poésie en prose, d'autres parlent de textes anecdotiques. Le corpus inclut ensuite les recueils de nouvelles : *Mondo et autres histoires*, *La ronde et autres faits divers* ainsi que *Printemps et autres saisons*. Les essais littéraires sont de différentes époques. *L'extase matérielle* et *L'inconnu sur la terre* traitent de thèmes généraux, tandis que *Trois villes saintes* et *Le rêve mexicain ou la pensée interrompue* s'intéressent exclusivement à la culture amérindienne. La culture amérindienne est également le principal intérêt des ouvrages à vocation ethnologique, *Les prophéties du Chilam Balam* et *La fête chantée*, tandis que *Sirandanes* s'intéresse à la culture de l'île Maurice. En outre, deux livres pour enfants sont inclus dans le corpus : *Voyage au pays des arbres* et *Pawana*. Enfin, sont présents *Diego et Frida*, la seule biographie, et *Gens des nuages*, le récit de voyage.

C'est premièrement à travers une étude sur la structure lexicale du corpus que nous pouvons observer l'influence de la riche variation typologique des textes.

3. LA STRUCTURE LEXICALE

Les différentes recherches sur la structure lexicale offrent la possibilité, indépendamment du contenu lexical, de situer, de distinguer et de comprendre la structure formelle des textes afin de pouvoir comparer différents discours, genres, époques ou auteurs différents au niveau exogène aussi bien qu'au niveau endogène, les parties de l'œuvre d'un écrivain ou de tout autre producteur de texte ou de parole.

Toutefois, dans les études statistiques, pour effectuer des analyses quantitatives différentes fines, riches en information, les fréquences absolues ne suffisent pas. Il est important de connaître l'étendue de son corpus et de ses parties. En effet, les valeurs de N et de V^1 ne sont pas liées par une relation fixe. Certes, V augmente quand N augmente, mais la relation qui les unit n'est pas une loi simple : l'étendue du vocabulaire est, comme l'exprime Charles Muller (1979) "un fait de style limité par des contraintes linguistiques".

Or, les calculs effectués par le logiciel Hyperbase, utilisé dans cette étude, permettent de mesurer l'étendue des textes dans le corpus en prenant en compte ces contraintes. Les calculs du poids relatif, c'est-à-dire l'espérance mathématique de l'événement : occurrence d'un mot dans le texte considéré (P) et non-occurrence de ce mot dans le même texte (Q=1-P), permettent l'emploi des lois classiques de la lexicométrie, principalement la

loi normale et la loi binomiale (Muller, 1977). Ces lois servent aux calculs de pondération dans les différents traitements statistiques.

Les graphiques suivants permettent de visualiser une des caractéristiques de notre corpus ; le premier histogramme regroupe les 100 plus hautes fréquences et rend compte de leur distribution, et le deuxième illustre la distribution des hapax[2] dans les différentes œuvres du corpus[3] :

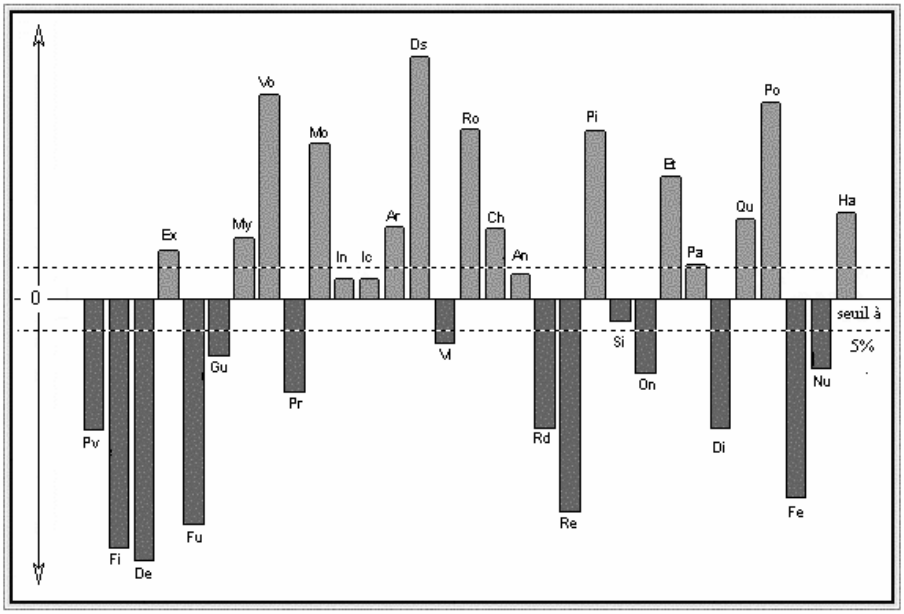


Figure n°1 : La distribution des plus hautes fréquences à travers le corpus.

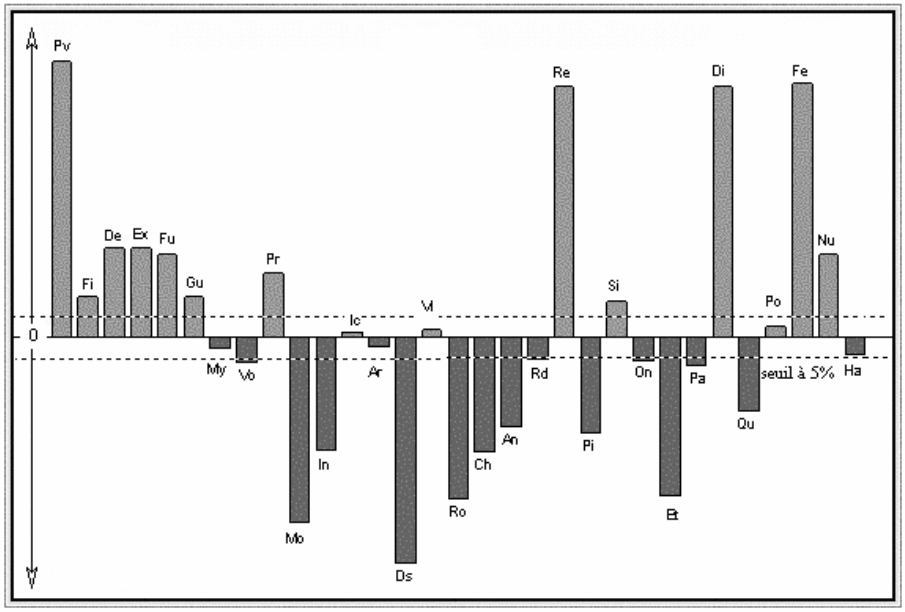


Figure n°2 : La distribution des hapax à travers le corpus.

Les histogrammes illustrent parfaitement l'inversion du mouvement dans la distribution de fréquences. Il est aisé de constater que les livres qui contiennent le plus d'hapax sont les plus "pauvres" en hautes fréquences. L'accueil fait aux hapax est en effet déterminé par le genre littéraire. Les taux négatifs, à quelques exceptions près, sont à trouver dans les romans et dans les nouvelles. Dans les autres ouvrages, ainsi que dans les œuvres, inspirés par l'école "nouveau roman", la quantité d'hapax est importante et, inversement, les hautes fréquences sont déficitaires.

L'étude la plus traditionnelle en lexicométrie est peut-être celle du rapport entre le nombre d'occurrences (N) et le nombre de vocables (V). Ce rapport donne une idée du nombre de mots différents comparé à l'étendue des textes et il permet, les valeurs correctement pondérées, de mesurer la richesse lexicale. L'analyse de la richesse lexicale des différents ouvrages reflète aussi souvent l'influence du genre dans lequel il s'inscrit. Notre corpus ne fait pas exception à cette règle, déjà bien documentée par ailleurs. En effet, les caractéristiques des différents genres se retrouvent dans notre corpus. Les romans et les nouvelles présentent le vocabulaire le plus "pauvre" tandis que les essais, les ouvrages ethnologiques et les récits de voyage offrent le vocabulaire le plus "riche". Dans ces derniers ouvrages, nous pouvons également noter la même tendance à la hausse de la richesse lexicale vers la fin de l'œuvre.

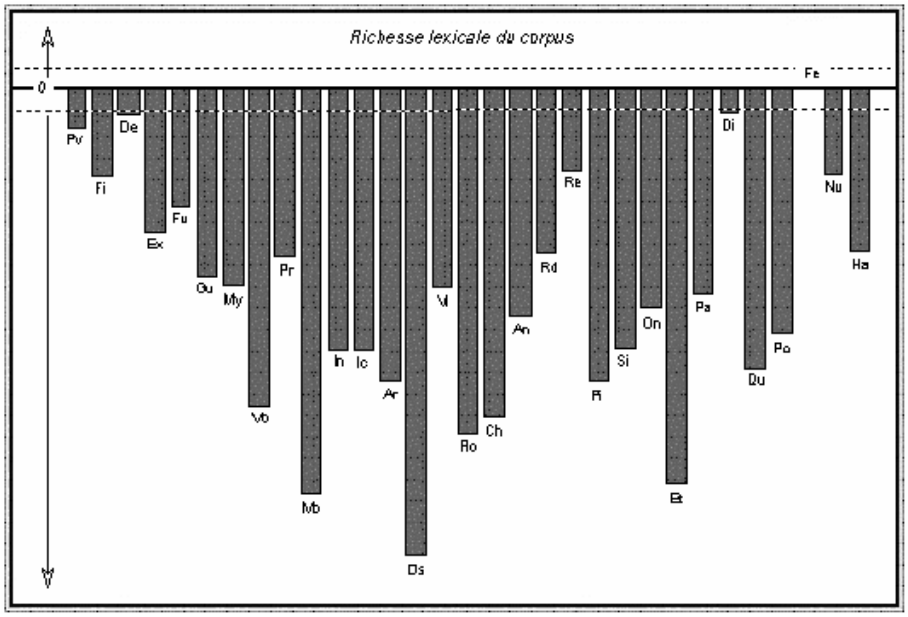


Figure n°3 : La richesse lexicale calculée sur l'étendue relative des textes, suivant la loi binomiale.

L'étude de l'accroissement lexical détermine l'apport du vocabulaire au fil du temps ; cet accroissement est, pour un segment déterminé du texte, le nombre d'unités nouvelles, c'est-à-dire n'ayant pas été employées antérieurement, qui apparaissent dans ce segment. Pour effectuer cette mesure, on découpe le corpus en tranches.

La représentation graphique ci-dessous rend compte de l'accroissement du vocabulaire dans l'ordre chronologique. Ici, le calcul fait appel à un ajustement des deux séries parallèles (vocabulaire cumulé et étendue cumulée) grâce à une fonction-puissance de type : $y = ax^b$ pour x = vocabulaire cumulé et y = étendue cumulée théorique. L'écart entre étendue théorique et étendue réelle est alors calculé pour chaque texte, puis pondéré par l'étendue de chaque texte :

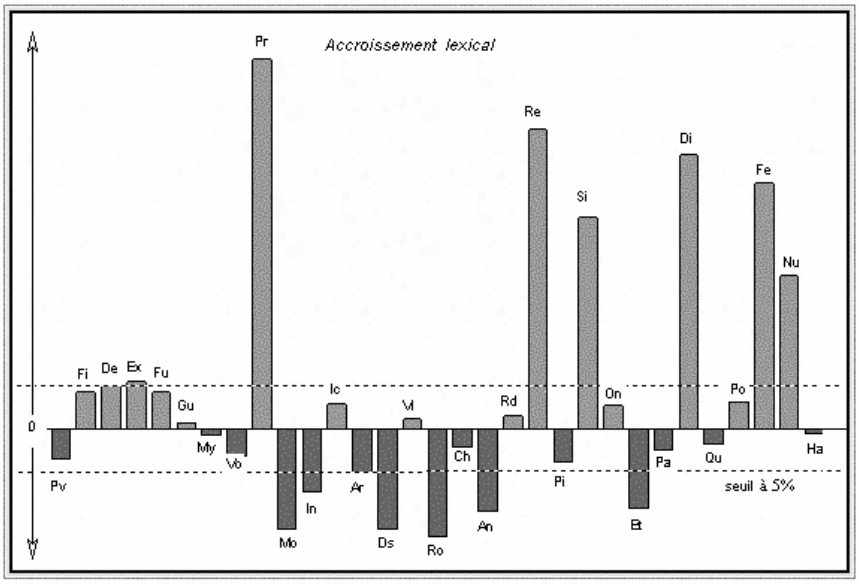


Figure n°4 : Accroissement lexical du corpus.

Le graphique qui, de gauche à droite, s'oriente selon la chronologie, nous permet de constater que les écarts autour de la moyenne, l'axe horizontal, sont de très grande ampleur, avec des ruptures et des reprises. Le seuil à 5 % est dépassé de nombreuses fois, avec des "pics" importants, dans le sens positif aussi bien que dans le sens négatif. L'étude de l'accroissement fait en effet très clairement apparaître, comme dans l'étude de la richesse lexicale et des hapax, l'opposition générique très importante du corpus : les œuvres qui contribuent le plus fortement à l'accroissement lexical sont les livres ethnologiques au sens large, alors que les romans et nouvelles "classiques" le ralentissent. En outre, un deuxième facteur se superpose au facteur générique : le facteur chronologique, qui divise Le Clézio en trois périodes principales. Nous avons pu constater que la courbe récurrente d'un vocabulaire croît de manière significative au début de l'œuvre et décline brusquement à partir de la fin des années 1970, pour s'accroître de nouveau vers la fin de l'œuvre sans que ces dernières valeurs atteignent les apports de la période initiale. La chute que nous avons observée dans nos différents histogrammes correspond bien à la rupture dans l'écriture de notre auteur, si souvent évoquée par les critiques littéraires. Enfin, la troisième période apporte des thèmes nouveaux à partir de 1987 sans pour autant présenter des apports lexicaux très importants, sauf quand le genre l'impose, comme dans les ouvrages ethnologiques, dans les essais et dans la biographie.

Globalement, l'analyse de la structure lexicale du corpus permet de constater, en premier lieu, le rôle très important du genre littéraire. Les essais, les ouvrages ethnologiques et la biographie présentent une richesse lexicale avec une grande spécialisation du vocabulaire, ainsi que des apports lexicaux importants dans notre corpus. En deuxième lieu, les différentes analyses mettent en évidence le facteur chronologique et l'évolution de l'œuvre. Les résultats confirment les intuitions contradictoires que peut avoir le lecteur de Le Clézio : d'un côté celle d'un vocabulaire riche, de l'autre celle d'un style pauvre, d'une écriture quelque peu répétitive. La bipolarité de la structure lexicale confirmée par l'analyse statistique, avec un vocabulaire qui tend soit vers l'abondance soit vers le dépouillement, est le fidèle témoin du paradoxe de l'écriture leclézienne et oppose ainsi le genre "nouveau roman" au genre "roman traditionnel".

Ces oppositions observées à l'étude de la structure lexicale d'un corpus ne sont pas indépendantes de la syntaxe. Par exemple, le déficit dans les basses et moyennes fréquences n'est pas un choix délibéré mais la conséquence d'un emploi intensif des mots grammaticaux – qui se concentrent dans les fréquences très élevées. C'est donc un choix syntaxique – dont nous relevons les effets dans le lexique. Il semble que Le Clézio fasse moins appel à un style recherché, au point de vue de la syntaxe, dans les ouvrages où il emploie beaucoup d'hapax, comme dans les ouvrages ethnologiques où la richesse d'hapax correspond souvent à la découverte d'une nouvelle culture. Inversement, dans les livres qui sont pauvres en hapax, comme dans les romans de la fin des années 1970, la richesse en hautes fréquences pourrait être un indice d'une plus grande complexité de la syntaxe.

L'opposition générique est en effet également à constater au niveau syntaxique, et l'analyse quantitative de la distribution des différentes parties du discours constitue une base impartiale et concrète pour permettre une étude syntaxique d'un corpus de taille.

4. LES PARTIES DU DISCOURS

La distribution des parties du discours dans les ouvrages littéraires n'est pas constante. Elle est fortement influencée par l'époque, l'auteur et le genre de discours. Et, l'emploi des catégories grammaticales dans un texte donné peut constituer un indice très révélateur. En effet cette distribution, qui est bien un critère de distinction des œuvres, manifeste peut-être des choix plus subtils que celui du vocabulaire – en tout cas moins liés à la thématique de chaque ouvrage, ce qui peut apporter à l'analyse des éléments nouveaux. Il s'agit en réalité de choix inconscients faits par l'auteur lors de la création et de l'élaboration d'un texte qui permettent au chercheur de distinguer des divisions grammaticales caractéristiques et personnelles.

Désormais, la quantification et la lemmatisation des corpus ouvrent la voie à cette composante essentielle de l'écriture. Cette analyse, qui demande l'accès à la forme canonique du mot, au lemme, ne peut guère se fonder sur la distribution des effectifs d'un corpus s'appuyant sur les formes graphiques. C'est la lemmatisation qui permet d'étiqueter le corpus selon les catégories grammaticales et de classer les éléments du vocabulaire selon leur appartenance à une catégorie spécifique. Les codes grammaticaux fournis par l'étiqueteur morphosyntaxique au cours de l'opération de lemmatisation "automatique" constituent ici un outil indispensable (Kastberg Sjöblom, 2002).

Le corpus "Le Clézio" a été traité avec la version d'Hyperbase lemmatisée selon le programme Cordial 7, qui aboutit au bout du traitement à quelque 200 codes grammaticaux différents, en utilisant toutes les combinaisons possibles. Nous en avons extrait les 11 catégories fondamentales parmi celles que propose le programme Cordial ; verbes, substantifs, adjectifs, déterminants, pronoms, numéraux, interjections, prépositions, adverbes, conjonctions et délimiteurs (signes de ponctuations). Pour une vision synthétique des accords qui lient les codes grammaticaux et les différents sous-corpus, nous avons recours à l'analyse factorielle de la liste de fréquences de ces différentes classes du corpus :

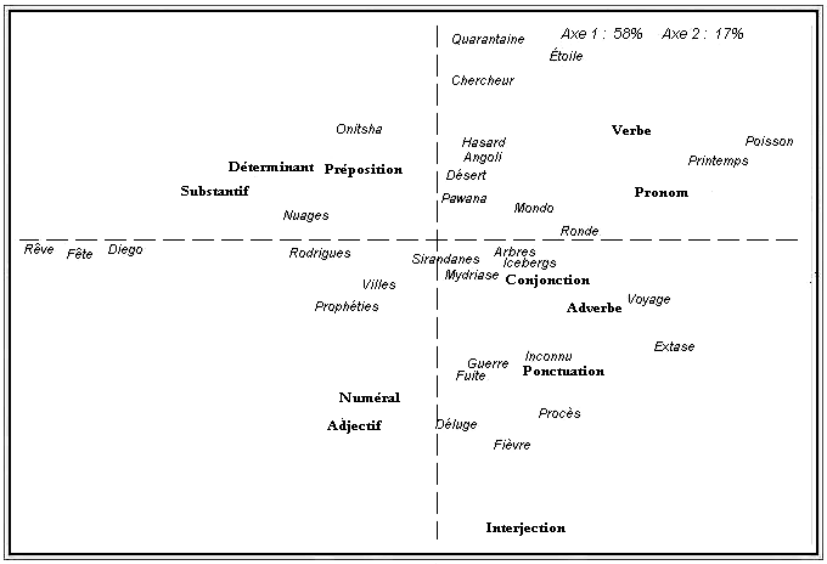


Figure n°5 : Analyse factorielle de la distribution grammaticale selon la lemmatisation par Cordial.

Nous voyons que le premier facteur oppose le long de l'axe horizontal la catégorie verbale à la catégorie nominale. Le substantif à gauche attire les

prépositions, les déterminants et les adjectifs, tandis que le verbe en haut à droite attire les pronoms et les adverbes.

Le second facteur parcourt la chronologie de l'écrivain du bas vers le haut du graphique. Les premiers ouvrages, *Le procès-verbal*, *La fièvre*, *Le déluge*, *La guerre* et *Le livre des fuites* se trouvent en bas du graphique autour des catégories secondaires qui témoignent d'une écriture foisonnante (adjectifs, adverbes et interjections). Les derniers romans, *La quarantaine*, *Poisson d'or*, *Etoile errante*, *Le chercheur d'or* et *Hasard*, se situent en haut du tableau autour des catégories fondamentales, témoignant peut-être d'un assagissement de l'écriture, d'un travail de simplification du style.

L'analyse factorielle rend ici également compte de l'opposition générique. Les ouvrages ethnologiques se regroupent à l'extrême gauche du graphique, les premiers romans appartenant à l'école du "nouveau roman" en bas à droite, tandis que les œuvres fictionnelles se trouvent au centre supérieur du tableau. Les ouvrages qui se trouvent au milieu sont les plus courts, tous genres confondus.

On observe souvent dans un corpus clos, comme nous venons de le faire, que deux camps, la catégorie nominale et la catégorie verbale, s'affrontent : la classe du verbe et les catégories qui lui sont proches (subordonnants, relatifs, pronoms et adverbes) s'opposent à la classe nominale qui réunit autour du substantif les adjectifs, les déterminants, les prépositions et souvent les coordinations. L'histogramme du quotient entre les 459.957 substantifs et les 321.108 verbes, se révèle en effet très sensible au genre[4].

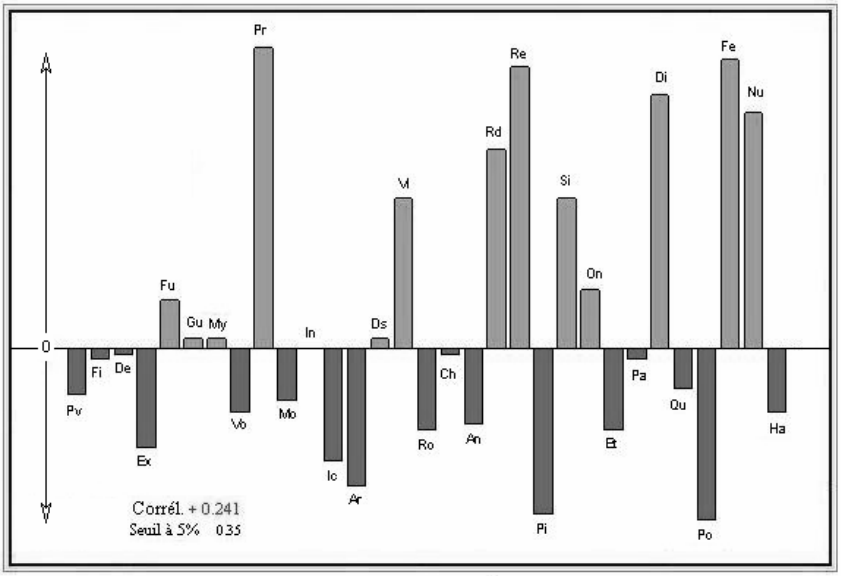


Figure n°6 : Histogramme du quotient substantifs/verbes.

Au début de la production de l'écrivain, dans sa période "nouveau roman", les deux courbes ne s'écartent point, elles se suivent au contraire, les deux catégories étant déficitaires dans cette partie de l'œuvre. C'est à partir de l'essai *L'extase matérielle* que l'opposition se déclare. Les deux courbes redeviennent parallèles dans *La guerre* et dans *Mydriase* pour se séparer de nouveau à partir de *Voyages de l'autre côté*. Dans les romans et dans les recueils de nouvelles qui paraissent entre 1975 et 1986, l'opposition des deux catégories est observable sans être très importante. Les écarts les plus importants − avec un fort déficit de verbes et un grand excédent de substantifs − sont à trouver dans les ouvrages d'ethnologie et dans les essais qui traitent du nouveau monde, comme *Le rêve mexicain* ainsi que dans la biographie *Diego et Frida*. *Poisson d'or* est le seul roman de cette époque qui présente un écart d'une grande amplitude, mais l'écart cette fois-ci témoigne d'un déficit important de substantifs et d'un excès de verbes.

Dans les œuvres non fictionnelles – les ouvrages ethnologiques, les essais, les récits de voyage et la biographie – l'évolution de l'opposition entre la catégorie du substantif et celle des verbes est en effet assez spectaculaire. Au début, les substantifs sont déficitaires et les verbes excédentaires (cf. *L'extase matérielle*), mais assez vite les rôles s'inversent et l'écart s'amplifie de façon importante. Il est difficile de fournir une explication précise, mais à un moment qui correspond à la découverte de la culture amérindienne et mexicaine, capitale pour notre écrivain, les substantifs commencent à abonder, tandis que les verbes diminuent de façon considérable. Cette découverte essentielle, Le Clézio veut en témoigner et il répète souvent : "Etre vivant c'est savoir regarder". Peut-être, à partir de ce moment, n'y a-t-il plus besoin du mouvement, des dialogues ni des verbes (d'action ou de parole), il suffit de regarder et de nommer : Le Clézio observe, décrit et partage ce qu'il voit avec ses lecteurs en recourant à de nombreux substantifs.

Cette bipolarité que nous pouvons observer entre les catégories des substantifs et des verbes chez Le Clézio n'a pourtant rien d'original : elle a été observée dans bien d'autres corpus. Étienne Brunet l'a bien remarquée dans ses diverses études et il souligne également le rôle important de l'opposition des genres littéraires (Brunet, 1985). De ce point de vue, l'œuvre de Le Clézio s'inscrit tout à fait dans la dynamique générale de la littérature française.

La tendance générale de la distribution des autres catégories grammaticales, les adjectifs, les verbes, et les adverbes etc., met presque toujours en évidence les mêmes phénomènes et les mêmes oppositions. Nous observons toujours en premier lieu l'opposition des genres littéraires ; les romans sont riches en verbes mais pauvres en substantifs tandis que les

biographies et les ouvrages d'ethnologies, par exemple, sont tous très riches en adjectifs et en substantifs, et pauvres en formes verbales. Ces typologies de texte, avec leurs nombreuses descriptions, sont nettement favorables à l'adjectif. Les essais, souvent d'un caractère poétique, le sont aussi, mais il s'agit là plutôt d'une coloration stylistique. En outre, apparaît généralement l'évolution chronologique de l'œuvre. L'adjectif notamment, après avoir été excédentaire au début de l'œuvre, régresse au fur et à mesure que l'œuvre progresse dans le temps et reflète ainsi un changement de style chez l'écrivain.

La distribution des verbes ne fait pas exception à cette tendance et à l'intérieur même de cette catégorie nous observons les mêmes phénomènes. Le logiciel Hyperbase permet désormais de distinguer et de regrouper les sous-catégories de verbes de façon automatique (Kastberg Sjöblom, 2002). L'analyse regroupe les verbes selon leur statut de principal ou d'auxiliaire, selon le mode, selon le temps exprimé ou bien selon la personne[5]

Les modes verbaux.

Les modes du verbe français se distinguent, selon la tradition et les textes officiels, en cinq ou en six classes : infinitif, participe, subjonctif, impératif, indicatif et conditionnel.

La distribution des différents modes dans notre corpus est celle que l'on trouve dans pratiquement tous les corpus littéraires, c'est-à-dire avec un indicatif qui domine largement (63,3%), et des participes et infinitifs qui occupent à peu près un quart du groupe (respectivement 19,2% et 14,4%). Quant aux autres modes, ils sont minoritaires (conditionnel 1,6%, impératif, 0,8% et subjonctif 0,7%).

L'analyse factorielle nous permet de situer les différents ouvrages de notre corpus par rapport à la distribution des modes :

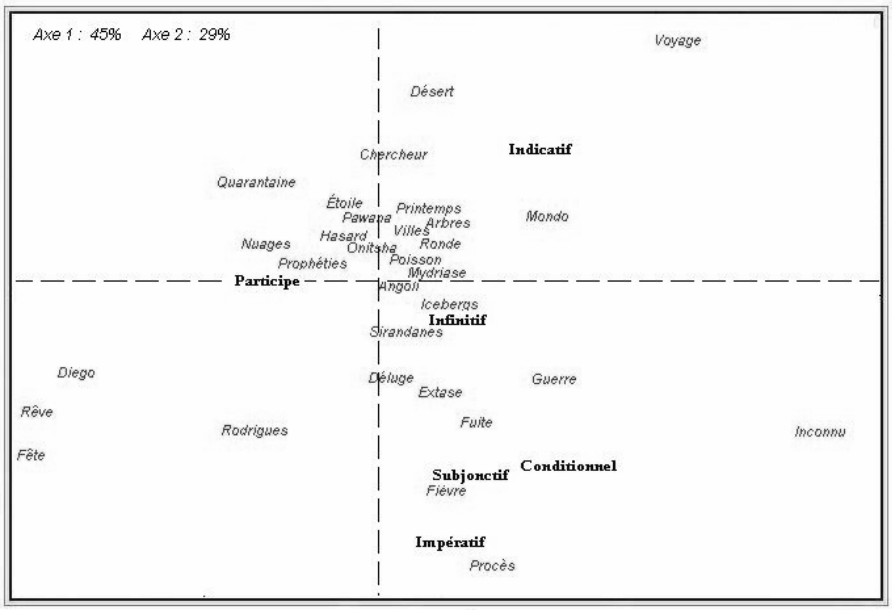


Figure n°8 : Analyse factorielle des modes verbaux dans le corpus Le Clézio.

Le premier axe du graphique met de nouveau en évidence l'opposition des genres littéraires, qui semble avoir une influence importante sur l'usage des différents modes verbaux. Les romans "traditionnels" se trouvent relativement au milieu du champ, dans la partie supérieure entre l'indicatif, qui est le mode traditionnel du récit, et le participe, qui est ici la trace des temps composés, en particulier celle du passé composé. Le subjonctif, le conditionnel et l'impératif appartiennent aux livres inspirés de l'école "nouveau roman", regroupés en bas et à droite du graphique. Les essais tardifs, *Le rêve mexicain* et *La fête chantée* ainsi que *Diego et Frida*, se trouvent ensemble éloignés du reste, en bas et à gauche. L'axe vertical reflète la chronologie de l'œuvre et rend bien compte de l'évolution dans l'écriture leclézienne. L'écriture souvent recherchée du début de l'œuvre – se manifestant par le recours à des modes comme le subjonctif, le conditionnel ou l'impératif – est abandonnée en faveur d'un style moins recherché, d'un récit plus traditionnel et d'une simplicité voulue par l'écrivain, privilégiant l'indicatif.

L'indicatif, mode du récit par excellence, domine effectivement le récit leclézien, mais la distribution interne montre toutefois des variations relativement importantes :

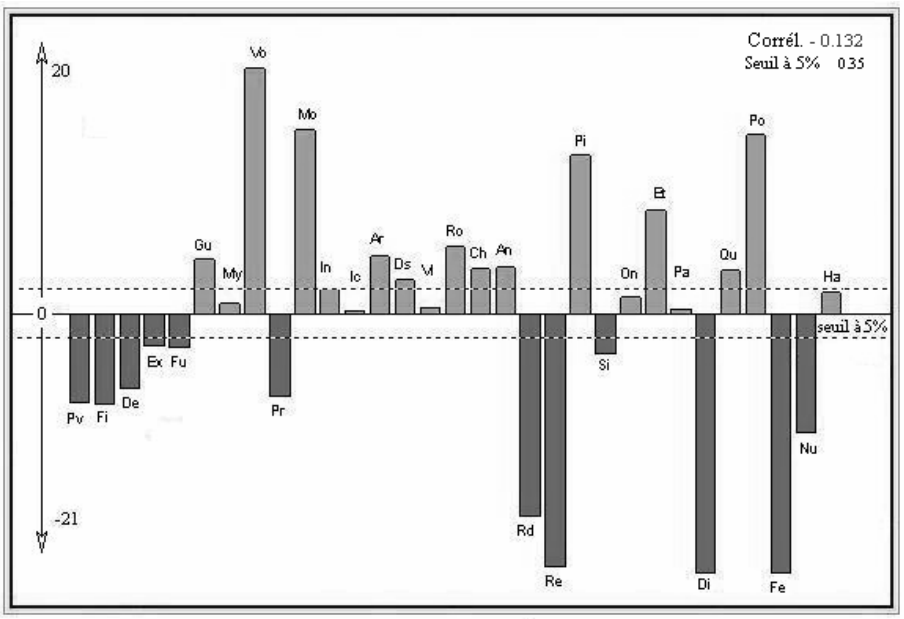


Figure n°9 : La distribution relative de l'indicatif dans le corpus (écarts réduits).

L'histogramme met en relief l'opposition des genres littéraires que nous avons constatée dans l'analyse factorielle. Les romans et les recueils de nouvelles privilégient l'indicatif – à l'exception des premiers romans et de *Voyages à Rodrigues* – tandis que ce mode est déficitaire dans les essais, dans les ouvrages d'ethnologie, dans la biographie et dans le récit de voyage. Nous pouvons aussi observer la tendance chronologique de cette distribution, avec des valeurs négatives de plus en plus importantes au fur et à mesure que l'œuvre progresse. C'est dans ces livres que nous avons pu observer auparavant les déficits les plus grands par rapport à la catégorie des verbes dans son ensemble.

L'étude des temps verbaux reflète également l'opposition générique dans notre corpus. L'analyse factorielle des temps de l'indicatif (présent, imparfait, passé simple et futur) rend compte des rapports qui lient les temps et les différents ouvrages du corpus :

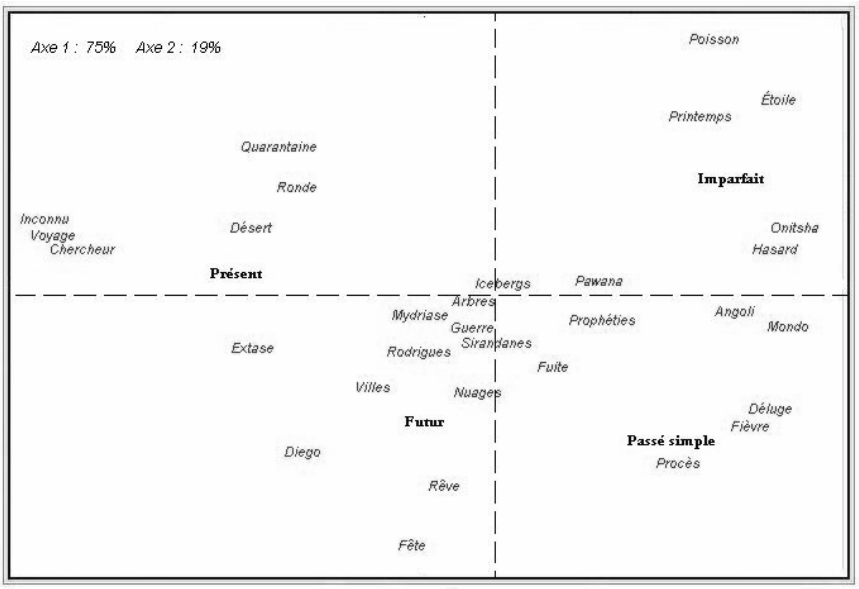


Figure n°10 : Analyse factorielle des temps verbaux.

Le premier axe de l'analyse oppose à nouveau les genres littéraires (bien que la division soit moins nette que dans les analyses précédentes) ; l'imparfait – du côté droit du graphique – est attiré par les romans de la deuxième période de Le Clézio. Le passé simple est employé dans la période "nouveau roman", les ouvrages ethnologiques semblent favoriser le futur. Quant au présent, sa position est plus difficile à expliquer, il semble que ce temps soit beaucoup employé par Le Clézio dans les romans de la deuxième période de l'œuvre. Le deuxième facteur de l'analyse factorielle rend compte de la temporalité et de l'évolution chronologique de l'emploi des différents temps verbaux chez Le Clézio. Nous trouvons les premiers livres en bas du graphique, une grande partie des ouvrages au milieu et tout en haut du tableau sont rassemblés les derniers romans : *Hasard*, *La quarantaine*, *Poisson d'or*, *Onitsha* et *Etoile errante*.

L'usage des différents temps verbaux dans un corpus est en effet un facteur qui – à part la fonction première : de nous situer dans le temps – est souvent déterminant pour le style d'un écrivain et change avec l'évolution d'une œuvre littéraire. Dans l'œuvre leclézienne, nous avons pu constater que l'emploi du verbe change au fur et à mesure que l'œuvre progresse, et que la fréquence des verbes est plus ou moins dominante selon l'époque ou les genres littéraires. Nous avons également vu qu'à l'intérieur de la catégorie verbale il y des variations importantes quant au mode aussi bien qu'à l'emploi des différents temps verbaux qui reflètent bien le changement perpétuel et la recherche de renouvellement de notre écrivain, tout en

gardant certaines constantes qui contribuent à donner au récit leclézien son caractère particulier, redondant, incantatoire et mystérieux.

Dans notre corpus, ce deuxième critère, morphologique, montre que la première période "nouveau roman" se démarque grammaticalement toujours du reste par son usage important du substantif et de l'adjectif, mais aussi par l'emploi de l'impératif et, paradoxalement pour une écriture expérimentale, par l'usage de formes temporelles très traditionnelles comme le passé simple. La rupture bien connue de l'œuvre leclézienne entraîne un changement vers une écriture qui privilégie l'action, et par conséquent les catégories verbales narratives, notamment les formes conjuguées à l'imparfait et les temps composés. L'étude des temps verbaux et de l'usage très personnel qu'en fait Le Clézio permet de mieux cerner une technique qui consiste à donner au récit cette valeur universelle, tant appréciée par ses lecteurs.

Une écriture qui change est une des caractéristiques fondamentales de notre corpus. En effet, il n'y a pas de "stabilisation" du style mais, au contraire, des écarts grandissants chez Le Clézio. Toutefois, bien que les procédés morphosyntaxiques ne soient pas statiques, que les techniques d'expression changent, qu'elles évoluent et qu'elles soient constamment mises en question, c'est l'opposition générique qui reste prépondérante.

Enfin, troisième critère d'analyse, l'étude du contenu du discours qui implique la signification des mots, les différentes catégories lexicales ainsi que l'étude des spécificités – positives et négatives – permet de dégager les caractéristiques thématiques d'une œuvre et de son évolution.

5. LE CONTENU DU DISCOURS

Aucun lecteur de Le Clézio n'est surpris par les résultats de l'étude statistique des thèmes de l'œuvre : certaines thématiques sont très importantes, la nature – terrestre et marine –, les couleurs ; d'autres comme le milieu urbain, les parties du corps, les insectes et le minuscule sont très présentes au début de l'œuvre, mais perdent de l'importance au fur et à mesure que l'œuvre progresse ; de plus leur importance varie selon le genre de texte. Dans la perspective endogène, les variations à l'intérieur du corpus sont toutefois importantes et l'étude statistique des spécificités permet de les cerner.

Un mot appartient au vocabulaire spécifique d'une partie d'un corpus lorsque sa fréquence relative dans ce texte, ou dans cette partie, s'écarte *significativement* de celle observée dans l'ensemble du corpus. Si la fréquence dans la partie étudiée est supérieure à la fréquence attendue, on dit que le mot est une *spécificité positive* ; dans le sens contraire, la spécificité est dite *négative*. En revanche, si l'écart n'est significatif dans aucune des parties du corpus, on dit que le mot est "non-spécifique" ou "commun". Si

les tests permettent de conclure avec plus ou moins de certitude à la spécificité d'un mot, la "non-spécificité" n'a pas le même statut : la normalité de la distribution est l'hypothèse la plus probable, mais elle n'est pas prouvée.

Le logiciel Hyperbase effectue ces calculs et permet l'observation du vocabulaire spécifique de chacune des 31 œuvres du corpus Le Clézio en s'appuyant sur les mêmes techniques déjà utilisées dans les analyses de structures et du rythme, à savoir la loi normale et les écarts réduits. Les spécificités propres au corpus peuvent aussi être repérées grâce à un système de segmentation interne du texte en plusieurs parties comparables[6]. Les résultats de l'analyse sont très nets, les listes de mots obtenues reflètent parfaitement le thème de l'ouvrage et nous donnent le profil caractéristique de chaque livre.

Dans l'étude de la distance lexicale, où il s'agit de considérer le vocabulaire intégral de chacun des textes du corpus et de repérer ceux qui partagent des thèmes semblables, nous avons trouvé, comme auparavant – dans les analyses structurelles et stylistiques du corpus –, des oppositions fortes entre les différents genres littéraires et un regroupement des livres appartenant à une même variété générique. La division tripartite à l'intérieur du genre romanesque écarte les ouvrages initiaux inspirés de l'école "nouveau roman" des autres – en indiquant que le changement thématique intervient déjà avec le roman *Voyages à l'autre côté* de 1975 – pour ensuite distinguer les autres ouvrages romanesques en un deuxième et un troisième regroupements du corpus qui tiennent bien compte de la chronologie de l'œuvre et de son évolution.

6. CONCLUSION

Ainsi, la numérisation et l'analyse lexicométrique de la quasi totalité des textes lecléziens nous ont permis de mettre en exergue l'importance de l'opposition générique qui s'observe à tous les niveaux de l'écriture : dans la structure, dans la syntaxe aussi bien que dans le vocabulaire. Ces résultats contredisent d'une certaine manière ce qu'a souvent écrit Le Clézio à propos de son écriture et des genres littéraires notamment dans l'ouvrage *La fièvre* (1965) :

> "Tout et rien. Je prenais des feuilles de papier, les plus grandes possible, et je les couvrais d'écriture, presque sans y prendre garde, presque au hasard. Mais ça n'avait aucun genre littéraire, c'était simplement de l'écriture."

En effet, le refus de genres est souvent une position idéaliste ou sociopolitique. Aussi, bien que Le Clézio refuse toute référence à un genre littéraire et que les critiques aient souvent souligné le mélange des genres dans un même ouvrage, nos analyses ont montré que l'appartenance à un

genre précis de chacun de ses livres est bien réelle. L'analyse statistique permet ici, par des étapes diverses, d'une part de confirmer de façon formelle et impartiale, les études antérieures "classiques", plus subjectives, d'un corpus littéraire ; d'autre part d'apporter des éclaircissements nouveaux, avec l'aide de ces outils informatiques qui permettent de prendre en considération simultanément l'intégralité de l'œuvre littéraire.

La distinction de typologies de textes n'opère, nous semble-t-il, ni à un niveau conscient lors de la production, ni à un niveau interprétatif. L'opération de classification par laquelle un lecteur donne une certaine cohésion à une suite textuelle est, certes, une opération de lecture-interprétation qui confère au discours une certaine structure compositionnelle, régi par la finalité des textes. Mais comme l'écrit Jean-Marie Adam (1992) on a parfois un peu trop tendance à séparer lecture et production :

> "La nature textuelle des faits de langue a pour conséquence la production d'un agencement de formes – une structure compositionnelle donnée – qui guide la lecture certes d'un énoncé achevé, mais qui guide aussi la première de toutes les lectures : celle opérée par le producteur lui-même au cours du processus de production de son discours. Avant de s'ouvrir sur l'espace interprétatif de toutes les lectures possibles, la nature séquentielle des faits de langue et l'existence de prototypes guident et matérialisent le processus interprétatif du producteur lui-même."

Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement au niveau pour ainsi dire "atomique", et cela transparaît très clairement dans les différents textes qui forment l'œuvre leclézienne.

7. RÉFÉRENCES

Adam J.-M. 1992. *Les textes : Types et prototypes*. Paris : Nathan. Coll. « fac. linguistique ».

Brunet E. 1985. *Le vocabulaire de Zola*. Paris-Genève : Champion-Slatkine.

Brunet E. 2001. *Hyperbase.* Manuel de référence, version 5.0. Nice : CNRS-InaLF, "Bases, corpus et langage" (UMR 6039).

Kastberg Sjöblom M. 2002. "Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus", in *JADT 2000, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, A. Morin, P. Sébillot (éds.). Saint-Malo : Irisa, Inria. (p. 391-402).

Kastberg Sjöblom M. 2002. *L'écriture de J.M.G. Le Clézio, une approche lexicométrique*. Nice : Université de Nice–Sophia Antipolis.

Le Clézio J.M.G. 1965. *La fièvre*. Paris : Gallimard. L'Imaginaire, n° 253.

Le Clézio J.M.G. 1969. *Le livre des fuites*. Paris : Gallimard. L'Imaginaire, n° 225.

Malrieu D. et Rastier F. 2002. "Genres et variations morphosyntaxiques", in *Actas del segundo seminario de la escuela interlatina de altos estudios en lingüística aplicada, Matemáticas y tratamiento de corpus, San Millán de la Cogolla, 19-*

23 septiembre de 2000, Angel Martin Municio (éd.). Logroño : Fundación San Millán de la Cogolla.
Muller CH. 1977. *Principes et méthodes de statistique lexicale*. Paris : Hachette.
Muller CH. 1979. "Calcul des probabilités et calcul d'un vocabulaire", Ch. Muller. *Langue française et linguistique quantitative*. Genève : Slatkine.
Rastier F. 1991. *Sémantique et recherches cognitives*. Paris : PUF, formes sémiotiques. (Réédition, PUF, 2001).
Rastier F. 2001. *Arts et Sciences du texte*. Paris : PUF, formes sémiotiques.

8. NOTES

(1) Le nombre d'occurrences total est représenté par N tandis que le V représente le nombre de vocables différents.
(2) Par le terme hapax, on désigne les vocables de fréquence 1 qui ont été rencontrés une seule fois dans un corpus, et conséquemment dans un seul texte.
(3) Pour l'interprétation du graphique la zone délimitée par les deux traits pointillés indique l'intervalle de fluctuation "normale", celle pour laquelle on ne peut pas écarter l'hypothèse que les fluctuations sont dues au hasard. Plus on s'écarte de cette zone, en dessous et en dessus, plus les écarts sont significatifs (du point de vue probabiliste).
(4) Le quotient est le rapport entre les deux séries. Il permet de voir comment se séparent les parallèles quand deux séries sont liées et parallèles. Comme les deux séries peuvent avoir un poids très inégal, la seconde est d'abord ramenée à la dimension de la première, proportionnellement, pour que le total des deux séries soit le même. Le quotient est calculé ensuite terme à terme, et s'équilibre nécessairement autour de la valeur 1.
(5) Il convient ici de souligner la difficulté des ambiguïtés et la relative opacité d'une analyse automatique qui ne permet pas à l'utilisateur de faire son choix et de trancher selon son propre jugement dans les cas ambigus.
(6) L'application fonctionne en accueillant des excédents de chaque partie du corpus par rapport aux autres, et la spécificité est déterminée par le calcul de l'écart réduit pour chaque forme dans chaque partie du corpus. Rappelons que ces différences de taille entre les unités de comparaison dans le corpus sont nivelées grâce à une pondération statistique.

LE LOGICIEL HYPERBASE : PRÉPARATION ET PRÉSENTATION DES CORPUS - APPLICATION À UN CORPUS DE LANGUE ANGLAISE

Vasilica Milea
Université de Metz

1. INTRODUCTION

Hyperbase, logiciel documentaire et statistique qui fonctionne à sa pleine capacité pour les textes en français (nous pensons particulièrement aux fonctions d'étiquetage et de lemmatisation) a attiré depuis sa création l'intérêt des chercheurs travaillant avec des corpus de langues diverses : le latin (Sylvie Mellet, Étienne Evrard), le portugais (Carlos Maciel, Tomas de Vilhena), mais aussi l'espagnol, l'italien, l'anglais. Les concepteurs du programme notent dans le guide d'utilisation du logiciel :

> Hyperbase s'applique à toute langue qui utilise l'alphabet latin, ce qui exclut notamment l'arabe, le cyrillique, le grec et les idéogrammes chinois […] Le français a cependant deux privilèges : les dialogues et les messages visibles à l'écran sont dans cette langue et la comparaison externe est faite avec les données du Trésor de la Langue Française (p. 3-4)

Depuis peu, des comparaisons avec un corpus de référence sont possibles pour l'anglais (le BNC, British National Corpus) et le portugais (corpus extrait du journal *Publico*). Ces fonctions sont intégrées au logiciel et rendent les résultats du travail d'autant plus prometteurs.

À la lumière de ces exemples et de ces fonctionnalités, nous nous sommes proposée de soumettre à l'analyse d'Hyperbase un corpus de langue anglaise. Les fonctions statistiques du logiciel et surtout la perspective d'avoir des résultats d'une grande finesse sont les autres raisons qui nous ont guidées dans nos choix méthodologiques. Notre démarche n'est pas la première, et donc elle n'est pas singulière dans son genre. Nous considérons néanmoins que les spécificités internes de l'anglais (comme de chaque langue) imposent une préparation et des précautions particulières que nous allons exposer.

Le corpus soumis à l'analyse avec le logiciel Hyperbase contient 67 nouvelles d'Edgar Allan Poe, écrites entre 1832 – 1849. Ce corpus, *Poe*, a été constitué en respectant certaines règles de taille (2 000 à 6 000 mots) et d'homogénéité (appartenance au même genre littéraire). Le contexte d'analyse critique ne fait pas défaut pour l'œuvre d'Edgar Allan Poe ; cependant, les études linguistiques de son œuvre et de son style n'abondent pas – celles de linguistique quantitative encore moins.

Notre recherche porte sur la subjectivité dans le langage et la place du sujet parlant dans le discours littéraire. La plupart des nouvelles qui constituent notre corpus sont écrites à la première personne, ce qui a guidé notre analyse des marqueurs de la subjectivité, à commencer par les pronoms personnels et la classe des adjectifs qualificatifs. Afin de valider et de renforcer les résultats de notre recherche sur ces marqueurs, nous avons constitué un corpus de référence, composé des œuvres de 40 écrivains, contemporains de Poe. La nouvelle américaine du XIX^{e} siècle est le genre privilégié de ce corpus, intitulé *EASS - Early American Short Story*.

2. PRÉSENTATION DES DONNÉES

2.1. Mise en forme et désambiguïsations graphiques

2.1.1. TRAITEMENTS PRELIMINAIRES. Nous ne nous attarderons pas sur l'origine des ressources textuelles, abondantes sur Internet, surtout pour les écrivains du XIX^{e} siècle, dont les œuvres ne sont plus soumises au copyright. Un balisage manuel des données s'impose comme premier traitement du corpus, afin d'éliminer les citations et les mots en alphabet grec, les éventuelles erreurs d'orthographe (résidus de la reconnaissance optique des caractères) ainsi que les passages chiffrés, tels :

> « 53‡‡†305))6*;4826)4‡)4‡;806*;48‡8¶60))85;1-(;:*8-83(88)5*‡
> ;46(;88*96*?;8)*‡(;485);5*†2:*‡(;4956*2(5*- 4)8¶8*;40692
> 85);)6†8)4;1(‡9;48081;8:8‡1;48†85;4)485†528806*81(‡9;48;
> (88;4(‡?34;48)4‡;161;:188;‡?; »

dans "Le scarabée d'or".

2.1.2. TRAITEMENT DES ITALIQUES. Le logiciel Hyperbase analyse les données textuelles présentées en fichier ASCII (texte seulement). Cette mise en forme spéciale fait disparaître les italiques, très nombreux dans les textes de Edgar Allan Poe :

> I well, too well, understood that to *think*, in my situation, was to be lost. (E. A. Poe, "The Imp of the Perverse")

Cette phrase sera mise en forme dans un fichier "texte seulement" de la façon suivante :

I well, too well, understood that to _think_, in my situation, was to be lost.

Hyperbase, à son tour, interprétera cette suite comme :

2 _
1 think

Chaque mot en italique introduit dans un fichier “texte seulement” deux tirets (_) qui peuvent fausser les résultats des analyses sur la ponctuation. Il est évident que les mots marqués par l’écriture en italiques portent les traces du sujet parlant ; c’est pour cette raison que nous avons pratiqué un marquage spécial de ces mots. La suppression des deux tirets et l’ajout de la lettre *q* en position finale :

think => thinkq

peut s’avérer utile au premier abord, si nous envisageons l’étude des mots marqués graphiquement comme porteurs de traces de la volonté du sujet parlant. Une simple recherche de concordance ou d’index des mots se terminant par *q* (lettre qui n’apparaît pas en position finale dans les mots anglais) produira la liste suivante :

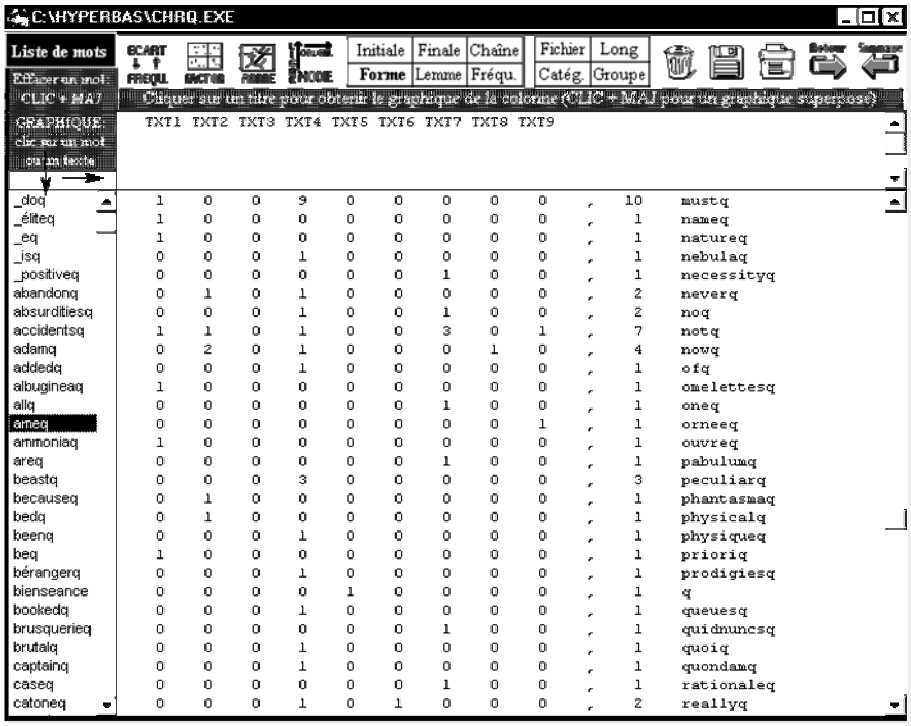

C:\HYPERBAS\CHRQ.EXE
Liste de mots
Initiale
Finale
Chaîne
Forme
Lemme
Fréqu.
Fichier
Long
Catég.
Groupe
TXT1 TXT2 TXT3 TXT4 TXT5 TXT6 TXT7 TXT8 TXT9
_doq 1 0 0 9 0 0 0 0 0 , 10 mustq
_éliteq 1 0 0 0 0 0 0 0 0 , 1 nameq
_eq 1 0 0 0 0 0 0 0 0 , 1 natureq
_isq 0 0 0 1 0 0 0 0 0 , 1 nebulaq
_positiveq 0 0 0 0 0 0 1 0 0 , 1 necessityq
abandonq 0 1 0 1 0 0 0 0 0 , 2 neverq
absurditiesq 0 0 0 1 0 0 1 0 0 , 2 noq
accidentsq 1 1 0 1 0 0 3 0 1 , 7 notq
adamq 0 2 0 1 0 0 0 1 0 , 4 nowq
addedq 0 0 0 1 0 0 0 0 0 , 1 ofq
albugineaq 1 0 0 0 0 0 0 0 0 , 1 omelettesq
allq 0 0 0 0 0 0 1 0 0 , 1 oneq
ameq 0 0 0 0 0 0 0 0 1 , 1 orneeq
ammoniaq 1 0 0 0 0 0 0 0 0 , 1 ouvreq
areq 0 0 0 0 0 0 1 0 0 , 1 pabulumq
beastq 0 0 0 3 0 0 0 0 0 , 3 peculiarq
becauseq 0 1 0 0 0 0 0 0 0 , 1 phantasmaq
bedq 0 1 0 0 0 0 0 0 0 , 1 physicalq
beenq 0 0 0 1 0 0 0 0 0 , 1 physiqueq
beq 1 0 0 0 0 0 0 0 0 , 1 prioriq
bérangerq 0 0 0 1 0 0 0 0 0 , 1 prodigiesq
bienseance 0 0 0 0 1 0 0 0 0 , 1 q
bookedq 0 0 0 1 0 0 0 0 0 , 1 queuesq
brusquerieq 0 0 0 0 0 0 1 0 0 , 1 quidnuncsq
brutalq 0 0 0 1 0 0 0 0 0 , 1 quoiq
captainq 0 0 0 1 0 0 0 0 0 , 1 quondamq
caseq 0 0 0 0 0 0 1 0 0 , 1 rationaleq
catoneq 0 0 0 1 0 1 0 0 0 , 2 reallyq

Figure 1 : Liste des mots en italique

De cette manière, nous avons pu mettre en évidence l'insistance de l'émetteur sur certains mots ayant une importante charge sémantique subjective comme :

> because, did, is, must, could, no, not, very

Ce type d'encodage ne peut pas être maintenu lors d'un deuxième encodage automatique (un mot comme *thinkq* ne serait pas reconnu par les lemmatiseurs ou les catégoriseurs), néanmoins, ces résultats préliminaires méritent toute notre attention.

2.1.3. TRAITEMENT DES MAJUSCULES. Une situation similaire de marquage graphique spécial que nous interprétons comme trace du sujet parlant ou de l'émetteur dans son texte apparaît dans l'emploi des majuscules. L'exemple suivant montre un nom commun et l'adjectif qui le précède qui sont écrits avec une majuscule :

> And this I did for seven long nights, every night just at midnight, but I found the eye always closed, and so it was impossible to do the work, for it was not the old man who vexed me but his **Evil Eye**. (E. A. Poe, "The Tell-Tale Heart")

Tout comme les italiques, les majuscules ponctuent le discours et fonctionnent comme de véritables pôles de signification dans le texte. Nous avons conservé les majuscules pour ces mots marqués graphiquement. Le traitement des noms propres et des majuscules avec Hyperbase nous a permis de mettre en évidence la prédilection de l'émetteur pour certains mots :

> bridge, civilize, easy, exchange, fated, literary, nature, revelation, seeing, snob, treason

et surtout une évidente prédilection pour la culture et la littérature européennes en général et celles françaises en particulier :

> Milton, Oedipus, Paris, Phoebus, Pindar, Prefecture, Prometheus, Pythagoreans, Rabelais, Ragoût, Rebus

2.1.4. CODES METALINGUISTIQUES. Notons également que le logiciel Hyperbase reconnaît les symboles :

> & et $

comme marques respectivement de titre et de page. Ces codes métalinguistiques ne doivent, par conséquent, exister dans le corpus que comme marques de segmentation, d'où la nécessité de les annoter lorsqu'ils apparaissent en tant que mots du texte. Dans un corpus d'anglais américain ces symboles apparaissent assez fréquemment pour que leur remplacement (par *and* et *dollar,* par exemple) soit justifié. Faute de cette simple opération, les risques de blocage et d'erreur lors du traitement ne sont pas négligeables.

2.2. Désambiguïsations grammaticales

2.2.1. TRAITEMENT DE LA FORME HOMOGRAPHE 'S. Après avoir reçu ce traitement de base, le texte est prêt pour l'étiquetage et la désambiguïsation grammaticale. Les formes ambiguës, telles *'s* imposent un traitement particulier, afin d'éviter les confusions entre *'s* marque du génitif et *'s* forme verbale dérivée de *be* ou de *have*. Ce qui plus est, un génitif comme *day's* sera lu par le logiciel comme :

1 day'
1 s

Un traitement simple comme l'élimination du *s* final peut sembler être la bonne solution, mais il suffit de prendre un exemple comme

It**'s** a work after my own heart.
(E. A. Poe, "A Tale of Jerusalem")

(dans lequel *'s* n'a rien d'un génitif, étant la forme contractée de *is*, inflexion du verbe *be*), pour comprendre que la seule solution est le repérage et l'encodage manuel, tâche extrêmement lourde pour notre corpus qui compte 300 000 mots.

Les mots qui contiennent une apostrophe présentent, de toute manière, une ambiguïté graphique qui reste problématique. Un mot comme *don't* sera analysé par Hyperbase comme :

1 don'
1 t

solution logique, mais qui multiplie les formes, si nous pensons aux formes *do* et *not* qui s'y ajoutent.

2.2.2. TRAITEMENT DES HOMOGRAPHES. Toutes les formes homographes des mots grammaticaux tels :

that, there, some

de même que les très nombreux homographes appartenant à de classes morphosyntaxiques différentes (nom-verbe ou nom-adjectif) présentent une ambiguïté que nous ne pouvons pas ignorer. Le logiciel Hyperbase fait un classement global de ces formes, intéressant, certes, mais qui reste grossier. L'unique solution pour désambiguïser le corpus est l'annotation.

2.3. Etiquetages grammaticaux systématiques

Annoter le texte manuellement est une tâche difficile et laborieuse quand il s'agit de grands corpus. Cette méthode peut s'avérer aussi moins fiable et moins cohérente que l'annotation automatique. Pour l'annotation du corpus, nous avons recours au logiciel CLAWS (Lancaster University, UCREL). L'étiquetage des catégories grammaticales est à notre avis plus productif que la lemmatisation, pour les objectifs de notre recherche, et pour une langue comme l'anglais qui présente peu d'inflexions. Le français a une

morphologie flexionnelle beaucoup plus riche que l'anglais, ce qui justifie un traitement de lemmatisation. Prenons comme exemple la catégorie des adjectifs, à laquelle nous nous intéressons tout particulièrement pour l'étude de la subjectivité :

> Le français déploie quatre formes flexionnelles pour la plupart de ses adjectifs : *bon – bonne – bons – bonnes*
>
> L'anglais, au contraire enregistre une forme unique : *good*

Les deux formes qui s'y ajoutent, le comparatif *better* et le superlatif *best* portent les traces du jugement subjectif du locuteur et méritent toute notre attention en tant que telles. Le logiciel CLAWS catégorise les degrés de comparaison des adjectifs, fonctionnalité plus appropriée pour notre analyse que la lemmatisation :

good	JJ
better	JJR
best	JJT

La forme *'s* sera étiquetée avec CLAWS, en fonction de son contexte, comme il suit :

GE	pour le génitif
VBZ	pour le verbe be
VHZ	pour le verbe have

Un mot comme *that* sera étiqueté de la façon suivante :

BCL	conjonction de subordination (dans les composés comme so that)
CST	conjonction de subordination
DD1	déterminant

En sollicitant les boutons "Initial" et "Final" des fonctions "Concordance" ou "Liste" d'Hyperbase, nous avons pu obtenir les index et les listes affinés de ce mot :

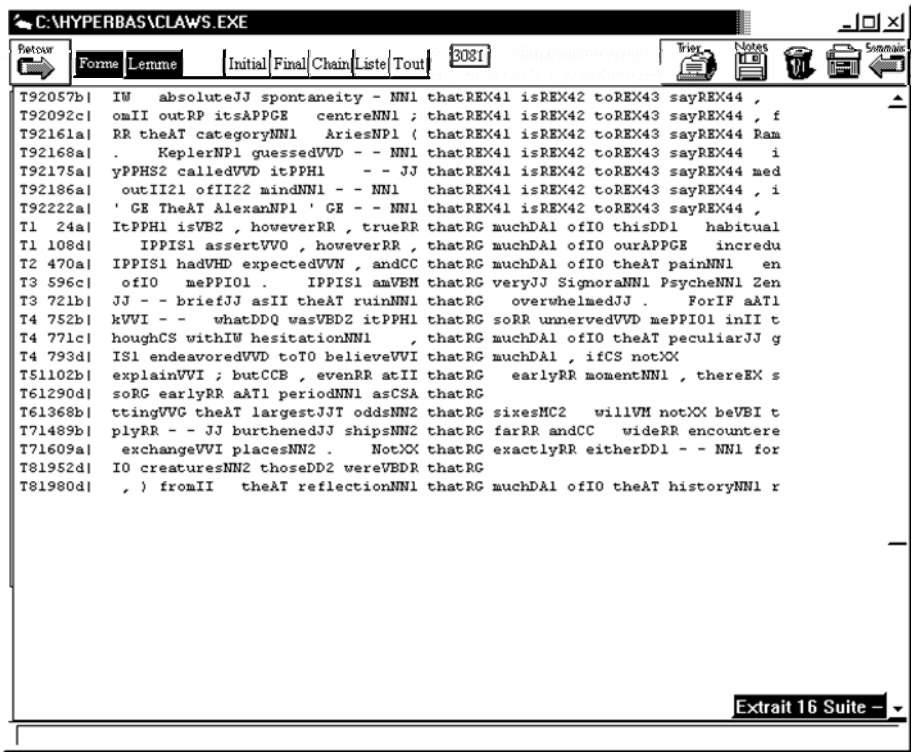

C:\HYPERBAS\CLAWS.EXE
Retour
Forme
Lemme
Initial
Final
Chain
Liste
Tout
3081
Trier
Notes
Sommaire
T92057b| IW absoluteJJ spontaneity - NN1 thatREX41 isREX42 toREX43 sayREX44 ,
T92092c| omII outRP itsAPPGE centreNN1 ; thatREX41 isREX42 toREX43 sayREX44 , f
T92161a| RR theAT categoryNN1 AriesNP1 (thatREX41 isREX42 toREX43 sayREX44 Ram
T92168a| . KeplerNP1 guessedVVD - - NN1 thatREX41 isREX42 toREX43 sayREX44 i
T92175a| yPPHS2 calledVVD itPPH1 - - JJ thatREX41 isREX42 toREX43 sayREX44 med
T92186a| outII21 ofII22 mindNN1 - - NN1 thatREX41 isREX42 toREX43 sayREX44 , i
T92222a| ' GE TheAT AlexanNP1 ' GE - - NN1 thatREX41 isREX42 toREX43 sayREX44 ,
T1 24a| ItPPH1 isVBZ , howeverRR , trueRR thatRG muchDA1 ofI0 thisDD1 habitual
T1 108d| IPPIS1 assertVV0 , howeverRR , thatRG muchDA1 ofI0 ourAPPGE incredu
T2 470a| IPPIS1 hadVHD expectedVVN , andCC thatRG muchDA1 ofI0 theAT painNN1 en
T3 596c| ofI0 mePPI01 . IPPIS1 amVBM thatRG veryJJ SigmoraNN1 PsycheNN1 Zen
T3 721b| JJ - - briefJJ asII theAT ruinNN1 thatRG overwhelmedJJ . ForIF aAT1
T4 752b| kVVI - - whatDDQ wasVBDZ itPPH1 thatRG soRR unnervedVVD mePPI01 inII t
T4 771c| houghCS withIW hesitationNN1 , thatRG muchDA1 ofI0 theAT peculiarJJ g
T4 793d| IS1 endeavoredVVD toT0 believeVVI thatRG muchDA1 , ifCS notXX
TS1102b| explainVVI ; butCCB , evenRR atII thatRG earlyRR momentNN1 , thereEX s
T61290d| soRG earlyRR aAT1 periodNN1 asCSA thatRG
T61368b| ttingVVG theAT largestJJT oddsNN2 thatRG sixesMC2 willVM notXX beVBI t
T71489b| plyRR - - JJ burthenedJJ shipsNN2 thatRG farRR andCC wideRR encountere
T71609a| exchangeVVI placesNN2 . NotXX thatRG exactlyRR eitherDD1 - - NN1 for
T81952d| I0 creaturesNN2 thoseDD2 wereVBDR thatRG
T81980d| ,) fromII theAT reflectionNN1 thatRG muchDA1 ofI0 theAT historyNN1 r
Extrait 16 Suite –

Figure 2 : Concordance (partielle) de that

L'adéquation du logiciel CLAWS pour le traitement de notre corpus ne vient pas seulement de ses fonctionnalités internes. Son degré de haute fiabilité est un facteur très important à prendre en compte, afin d'assurer un traitement complet, stable et univoque du corpus. C'est pour cette raison que nous avons préféré l'encodage automatique à un encodage manuel partiel, de certaines classes de mots. À ce stade de notre travail nous considérons que le corpus est prêt à l'emploi, à l'analyse avec Hyperbase.

3. TAILLE DU CORPUS, DÉCOUPAGES ET NOMBRE DE TEXTES

Pour ce qui est de la taille du corpus et du nombre de textes, les concepteurs du logiciel Hyperbase notent : "Dans sa version actuelle le programme accepte 81 textes. La longueur de chacun des textes n'importe guère." (Brunet, 1999) Notre corpus *Poe* divisé en 67 textes (295 839 mots) a été analysé par le logiciel en moins de deux heures. Bien que très pratiques pour la navigation dans le corpus, les divisions trop nombreuses rendent difficile

la lecture des résultats représentés graphiquement. Prenons comme exemple le graphique de la distribution du pronom personnel *I* dans le corpus :

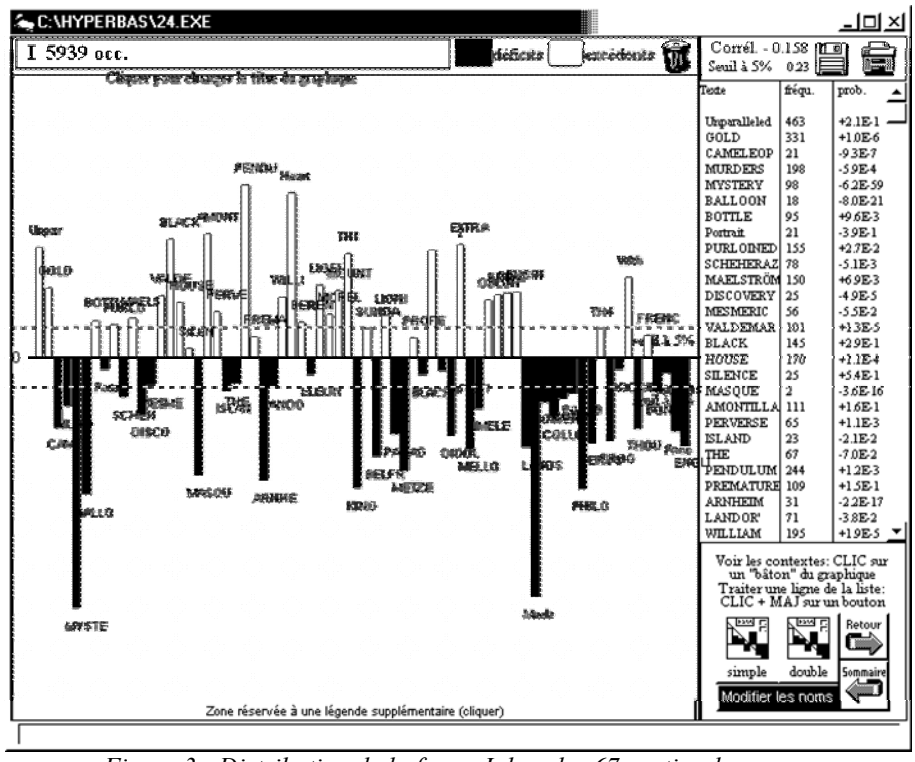


Figure 3 : Distribution de la forme I dans les 67 parties du corpus

Le même corpus sans aucune marque de segmentation (telles & ou $) sera découpé automatiquement par le logiciel en neuf parties. Cela rend la lecture plus aisée, mais le repérage sera moins précis au niveau de chaque texte :

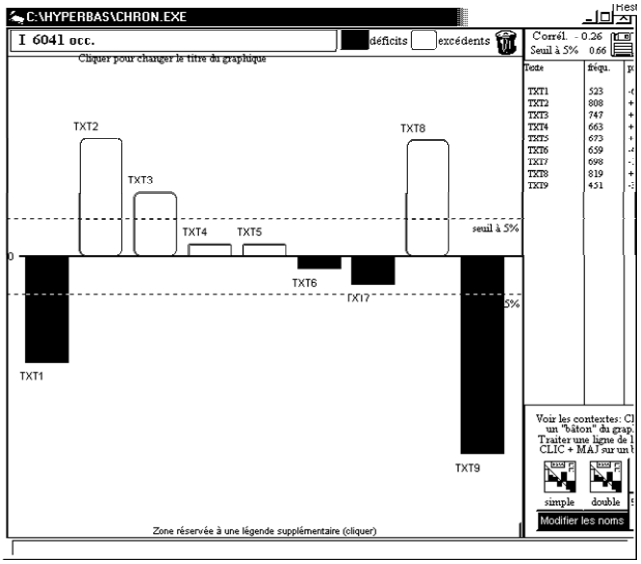


Figure 4 : Distribution de la forme I dans les 9 parties du corpus

Afin de pouvoir exploiter les fonctions d'Hyperbase qui décèlent l'évolution du lexique et la distribution des fréquences, nous avons disposé les textes dans le corpus par ordre chronologique. La lecture des résultats de l'analyse des distances lexicales se trouve ainsi facilitée par l'organisation chronologique des textes :

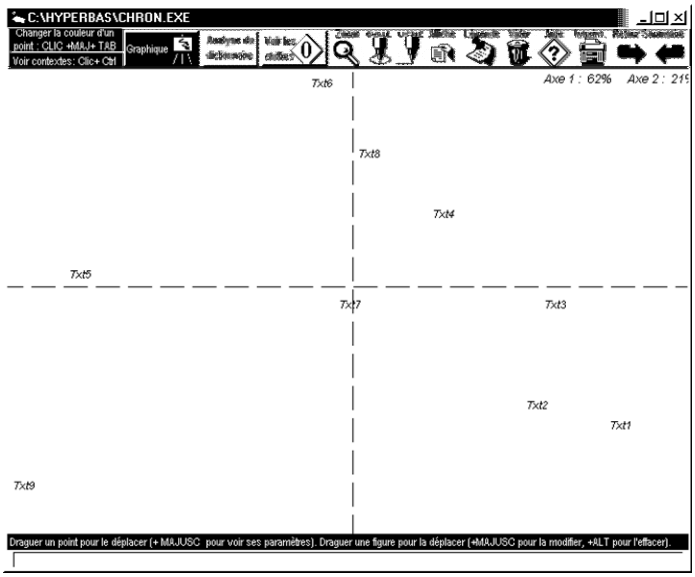


Figure 5 : Analyse factorielle de la distance lexicale

Comme nous pouvons le constater, les tests statistiques démontrent une évolution au niveau du vocabulaire : la distance qui sépare Txt1 (les textes écrits de 1832 à 1834) et Txt9 (les textes écrits de 1845 à 1849) en est la preuve. L'évolution constatée par les littéraires devient ainsi un fait prouvé statistiquement.

4. CONCLUSION

En conclusion, nous aimerions souligner l'intérêt des fonctions statistiques (calcul des corrélations, distribution des fréquences, analyse factorielle) qui sont les atouts incontestables d'Hyperbase.

Cependant, il ne faut pas perdre de vue l'importance primordiale de la bonne préparation du corpus ; la fiabilité des résultats en dépend directement. Bien que le logiciel puisse traiter les corpus à l'état brut, l'analyse gagne en qualité et précision lorsqu'elle s'exerce sur un corpus annoté. Plus importante que l'annotation est, à notre avis, la mise en forme du corpus et sa taille. L'absence d'encodage est préférable à un encodage métalinguistique défectueux qui peut très facilement bloquer le programme. Cette observation est valable pour tous les corpus. En ce qui concerne les corpus d'anglais, une version pour les données étiquetées, voire un catégoriseur intégré au logiciel améliorerait considérablement ce travail qui peut paraître, pour le moment, voué au hasard des formes.

5. RÉFÉRENCES

Brunet E. 2001.*Hyperbase*.**.** Université de Nice. CD-ROM. *Claws*. UCREL CLAWS7 Tagset, Lancaster University.

Brunet, E. 1999. *Hyperbase : Manuel de référence,* Version 4.0 (août 1999) pour Mac et Windows.

Habert, B., Nazarenko, A., Salem, A. 1997. *Les linguistiques de* corpus. Paris : Armand Colin.

Muller, C. 1992. *Initiation aux méthodes de la statistique linguistique*. Paris : Champion.

APPROCHE LEXICOMÉTRIQUE DE CORPUS D'INTERACTIONS VERBALES ENTRE UN ADULTE ET UN ENFANT EN COURS D'ACQUISITION DU LANGAGE. RÉSULTATS D'EXPÉRIENCE.

Luiggi Sansonetti
Université Paris III − EA 2290 SYLED − EA 170I CALIPSO, ILPGA

1. PRÉSENTATION DU CORPUS

1.1. Nature des corpus

Les corpus utilisés pour cette étude ont été recueillis par une étudiante dans le cadre de son mémoire de Maîtrise (Tissier, 2001). Le 1er corpus longitudinal, nommé « Julien », est composé de trois dialogues entre un adulte et un enfant autour d'un livre illustré. Cette étude diachronique présente le langage de Julien entre 5 ans 10 mois et 6 ans 4 mois, les dialogues étant espacés de trois mois. Le 2ème corpus longitudinal, nommé « Mathilde », est composé de trois dialogues entre le même adulte et un autre enfant, autour du même livre illustré. Cette étude présente le langage de Mathilde entre 4 ans 9 mois et 4 ans 11 mois, les dialogues étant espacés d'un mois.

1.2. Recueil des données

Pour ces corpus, l'adulte a utilisé un livre illustré[1] pour guider la discussion avec l'enfant. Dans un premier temps, l'adulte lit plusieurs fois l'histoire à l'enfant. Ensuite, pour l'enregistrement, l'adulte demande à l'enfant de raconter cette histoire. Enfin, après avoir recueilli les corpus, une fiche de situation est établie pour chaque enregistrement et les dialogues sont réécrits suivant des conventions de transcription adaptées aux objectifs de recherche de la linguistique de l'acquisition du langage.

1.3. Objectifs de recherche

La linguistique de l'acquisition recherche les processus d'élaboration du fonctionnement cognitivo-langagier, en observant tout particulièrement la syntaxe dans les énoncés de l'enfant, à partir d'interactions verbales avec un adulte en situations de dialogues spontanés. Cette approche s'appuie sur les avancées du courant interactionniste (Kerbrat-Orecchioni, 1999), ainsi que sur les recherches sur le français parlé et écrit (Blanche-Benveniste, 1997), qui éclairent l'analyse du langage adressé à l'enfant et le fonctionnement des interactions verbales. En effet, ces processus d'élaboration sont recueillis « en contexte authentique de production langagière, dans des situations de vie quotidienne, au cours de conversations familières » (Lentin, 1998). Le choix de la langue parlée, par rapport à la langue écrite, est justifié par les raisons suivantes :

- l'enfant apprend à parler dans des activités langagières dialogiques avec un adulte ;
- l'interaction doit avoir lieu dans une situation de parole spontanée pour garantir l'authenticité des énoncés ;
- et comme le souligne Blanche-Benveniste : « on ne peut pas étudier l'oral **par l'oral**, en se fiant à la mémoire qu'on en garde. On ne peut pas, sans le secours de la représentation visuelle, parcourir l'oral en tous sens et en comparer des morceaux » (1997).

Dans le cadre de travail du GARS[2] avec Blanche-Benveniste, comme du CRALOE[3] avec Lentin, le système adopté est la transcription orthographique de façon à tenir un juste équilibre entre la « fidélité » de ce qui a été dit et la « lisibilité » de la transposition par écrit (Ringeard et Lorscheider, 1977).

1.4. Outils TAL

Pour cette analyse automatique, nous utilisons principalement deux logiciels : Lexico3[4], et Cordial[5].

Lexico3 est un logiciel rassemblant des outils de statistiques textuelles et de lexicométrie élaborés dans le cadre de l'équipe « Lexicométrie et textes politiques », UMR 9952 (CNRS / École Normale Supérieur de Fontenay-St Cloud), puis dans l'équipe « LEXICO » de l'UPRES SYLED (Systèmes Linguistiques Enonciation et Discours – EA 2290 – Université de la Sorbonne Nouvelle Paris 3), par A. Salem, W. Martinez, C. Lamalle et S. Fleury.

Cordial est un correcteur orthographique et étiqueteur automatique développé par Synapse Developpement, permettant de procéder à des vérifications de la syntaxe du texte et d'associer aux mots la catégorie grammaticale correspondante, ainsi que d'autres renseignements (forme lemmatisée, genre…).

1.5. Informatisation du corpus

À chaque transcription sont liées des conventions de transcription, et à chaque analyse des représentations. Cependant, les logiciels ont eux aussi leur propre format de représentation de données. Lexico3, par exemple, n'accepte que le format « texte seul » (.txt). Cordial accepte, quant à lui, les formats « texte seul », « document Word » (.doc), et d'autres propres à l'application (« fichier étiqueté » (.cnr)…).

Nous devons donc, pour analyser les documents, les préparer au traitement, en nous donnant des normes de dépouillement, ce que nous préférons appeler « phases de préparation » : « la notion de **norme de dépouillement** doit être ici comprise comme une exigence de standardisation provisoire des textes contenus dans un corpus. Cette standardisation est destinée avant tout à les rendre comparables, à les stabiliser le temps d'une expérience » (Habert, Nazarenko et Salem, 1997).

Ces phases de préparation sont nécessaires pour rendre conforme le corpus aux exigences du logiciel Lexico3. D'autres sont motivées par un choix théorique concernant les unités à analyser et certaines, propres au logiciel, permettent de mettre en relief d'autres phénomènes que ceux recherchés dans le cadre de la linguistique de l'acquisition du langage.

Avant tout, pour passer le corpus à Lexico3, nous rassemblons les trois dialogues de Julien au format « document » en un seul corpus au format « texte seulement ». Nous faisons de même pour les dialogues de Mathilde. Ensuite nous procédons à différentes phases successives du traitement.

1.5.1. PHASE DE NETTOYAGE. Nous nettoyons le corpus en retirant tout ce qui n'est pas utile pour une analyse syntaxique (la présentation du dialogue, les commentaires, les indications phonétiques, les « ind. »). En effet, la fiche « situation » qui présente chaque corpus n'est pas utile pour l'analyse lexicométrique. Ensuite, il y a, de la part du transcripteur, des indications concernant le contexte de l'interaction (le micro tombe, l'enfant tourne la page, l'enfant montre l'image…). Au cours de la transcription, le transcripteur peut mettre des symboles phonétiques qui ne sont pas utiles pour les constructions syntaxiques.

Les marques de l'oral comme l'hésitation, notées « / », sont conservées. Ces marques peuvent mettre en relief une recherche, ou une tentative de construction syntaxique de la part de l'enfant, ou même montrer un échec de construction.

Malgré les progrès techniques des enregistreurs, il y a parfois des séquences inaudibles. Le transcripteur met alors l'indication « ind. » pour « indistinct ». Ces marques ne sont pas importantes pour notre analyse.

Enfin, nous retirons les marques de chevauchements de paroles. Ce qui est énoncé simultanément est présenté avec des espaces inutiles et parfois un objet graphique comme les accolades.

1.5.2. PHASE DE REMPLACEMENT. Nous procédons au remplacement des doutes par l'estimation. Quand le transcripteur n'est pas sûr de la « traduction », il y a la forme entendue suivie de ce qu'elle pourrait signifier : ce qui est noté « pa(r)ce que zaime (=j'aime) » se réécrit, dans un premier temps, « pa(r)ce que j'aime ».

De plus, dans le cas de mots où certaines lettres ou syllabes ne sont pas prononcées, le transcripteur les fait apparaître entre parenthèses pour la compréhension du corpus : pa(rce) que. Nous retirons aussi les « ( ) » pour réécrire le mot entièrement (« ouv(r)e » se réécrit « ouvre »). En effet, Lexico3 constitue un dictionnaire de formes utilisées. Et, dans notre cas, nous aurions compté deux formes différentes : « ouvre » et « ouv », alors qu'il s'agit de la même.

Enfin, toute majuscule est remplacée par sa minuscule pour ne pas avoir deux formes distinctes alors que c'est la même (sauf pour les noms propres où la majuscule est conservée).

1.5.3. PHASE DE BALISAGE. Les clés notées entre chevrons « <xxx> » peuvent partitionner le corpus en dialogues (1, 2, 3) et en locuteurs (adulte, enfant). Ce qui permet de visualiser l'évolution de l'emploi de la forme cherchée sur l'ensemble du corpus, ou sur l'ensemble des énoncés de l'un ou de l'autre locuteur :

- balisage des dialogues : <dial=ju1> ;
- balisage des locuteurs : <ad=…>, <ju=…> ;
- balisage de la référence de l'énoncé : <…=101> correspond à <…=dialogue n° 1 énoncé n° 01>.

De plus, il est possible d'ajouter des signes spécifiques pour segmenter autrement les énoncés :

- balisage des énoncés avec le délimiteur des énoncés « § » qui permet de revenir au texte plus facilement. Un carré représentera un énoncé.

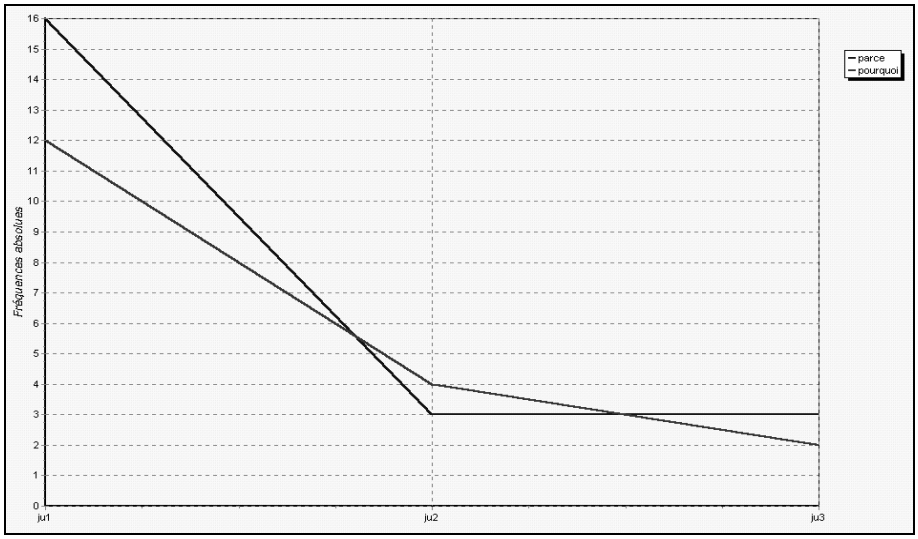
1.5.4. PHASE DE PRÉSENTATION. Les énoncés sont regroupés par 50 pour une question de présentation : nous aurons sur une ligne les énoncés de l'adulte et juste en dessous ceux de l'enfant, pour mieux repérer, par exemple, les reprises et reformulations.

2. QUELQUES ANALYSES

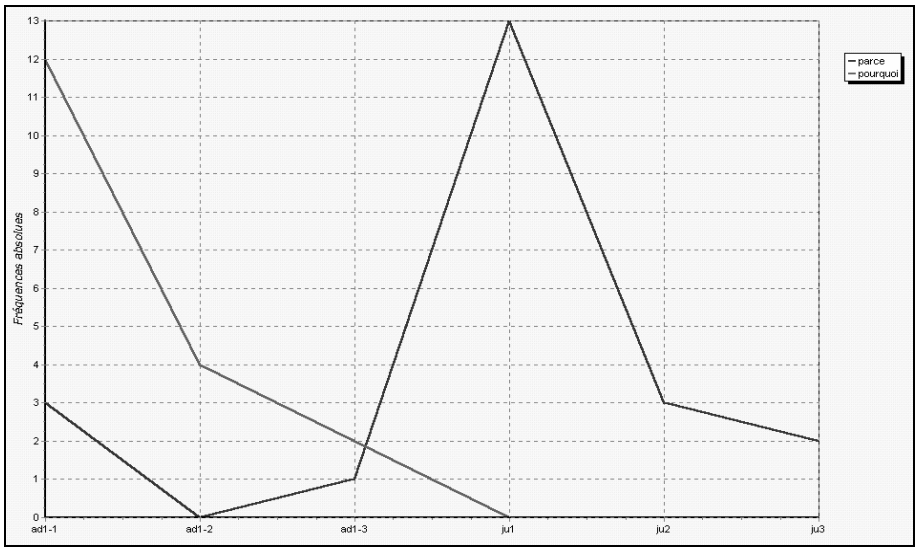
2.1. Corpus de Julien

Une fois le corpus informatisé, nous le passons à Lexico3 pour le segmenter et obtenir le dictionnaire des formes classées par fréquence ou par ordre alphabétique. C'est à partir de cette liste des formes que nous procédons à l'analyse proprement dite.

Nous commençons notre recherche sur « parce que » qui se trouve être presque toujours, surtout à l'oral, en construction syntaxique incomplète du fait de l'absence de la proposition principale. Nous vérifions en même temps s'il s'agit systématiquement d'une réponse à une question de type « pourquoi ». Nous obtenons alors un lien très étroit entre ces deux formes (Graphiques 1-2).



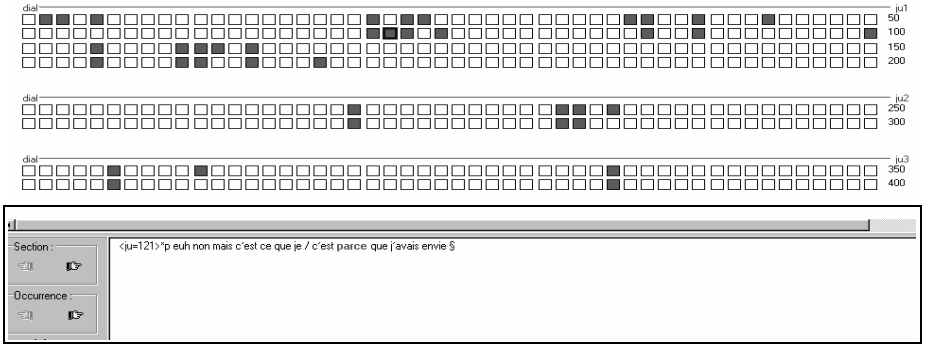
Graphique 1 : Répartition par dialogues



Graphique 2 : Répartition par locuteurs

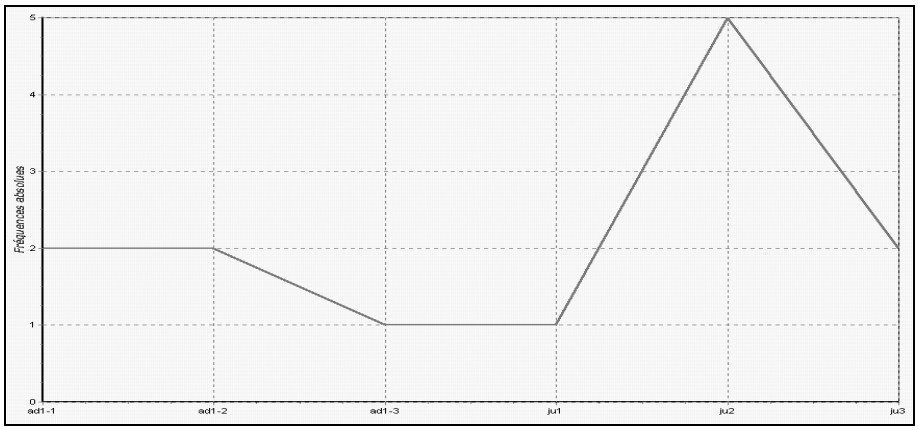
À l'aide de ce graphique, nous remarquons que l'enfant ne prononce pas de « pourquoi ».

Avec la carte des paragraphes nous pouvons revenir au texte et vérifier que le « parce que » de l'enfant fait systématiquement suite, immédiatement ou non, à une question de la part de l'adulte (Graphique 3).

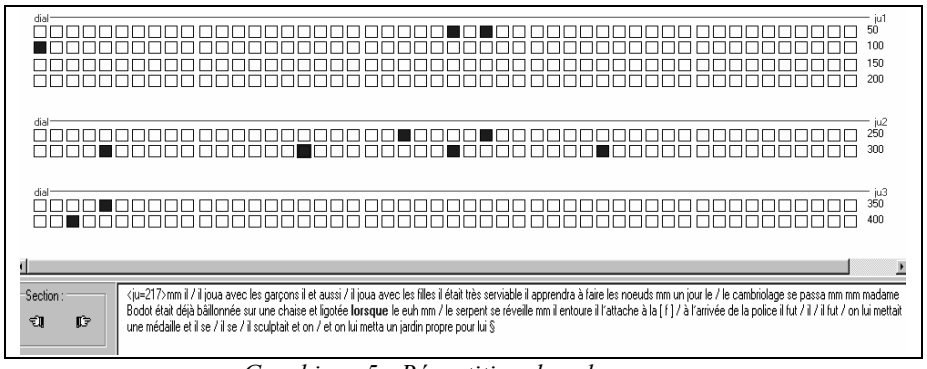


Graphique 3 : Répartition dans le corpus

Nous pouvons procéder à diverses recherches sur l'emploi de telle ou telle forme, de même qu'à un groupe de formes. Par exemple, nous recherchons les emplois des formes « quand, pendant, lorsque » (Graphiques 4-5) :



Graphique 4 : Répartition par locuteurs

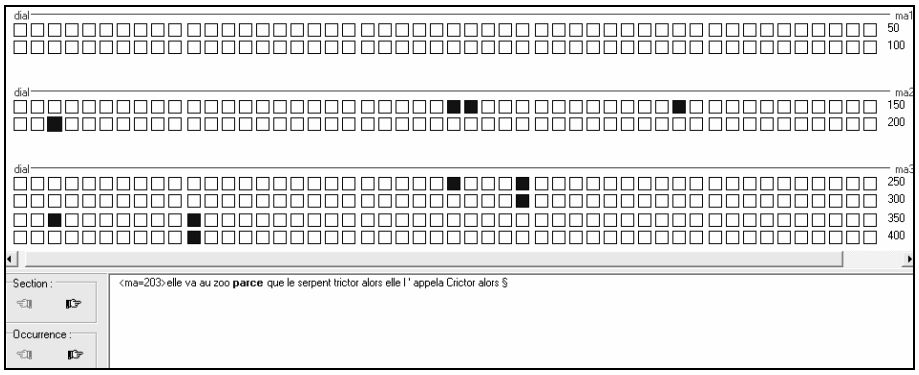

dial
ju1
50
100
150
200
dial
ju2
250
300
dial
ju3
350
400
Section :
<ju=217> mm il / il joua avec les garçons il et aussi / il joua avec les filles il était très serviable il apprendra à faire les noeuds mm un jour le / le cambriolage se passa mm mm madame Bodot était déjà bâillonnée sur une chaise et ligotée **lorsque** le euh mm / le serpent se réveille mm il entoure il l'attache à la [!] / à l'arrivée de la police il fut / il / il fut / on lui mettait une médaille et il se / il se / il sculptait et on / et on lui metta un jardin propre pour lui §

Graphique 5 : Répartition dans le corpus

Nous remarquons par cette carte que les emplois de ces formes par l'enfant ne sont pas en réponse aux propositions de l'adulte, et qu'il s'agit d'emplois spontanés dus à la situation de narration ou de récit d'histoire.

2.2. Corpus de Mathilde

Nous procédons aux mêmes types de recherches sur le corpus de Mathilde, et nous remarquons que ni l'adulte ni l'enfant ne posent de question « pourquoi ». Les « parce que » énoncés par l'enfant, comme par l'adulte, sont plusieurs fois réalisés dans une construction syntaxique complète. Pour le premier « parce que » de l'enfant, il se trouve en essai de construction, puisqu'il y a abandon de la structure (Graphique 6).



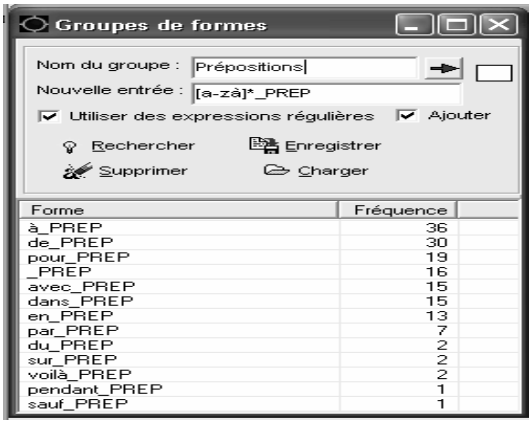
Graphique 6 : Répartition par locuteurs

2.3. Corpus étiquetés

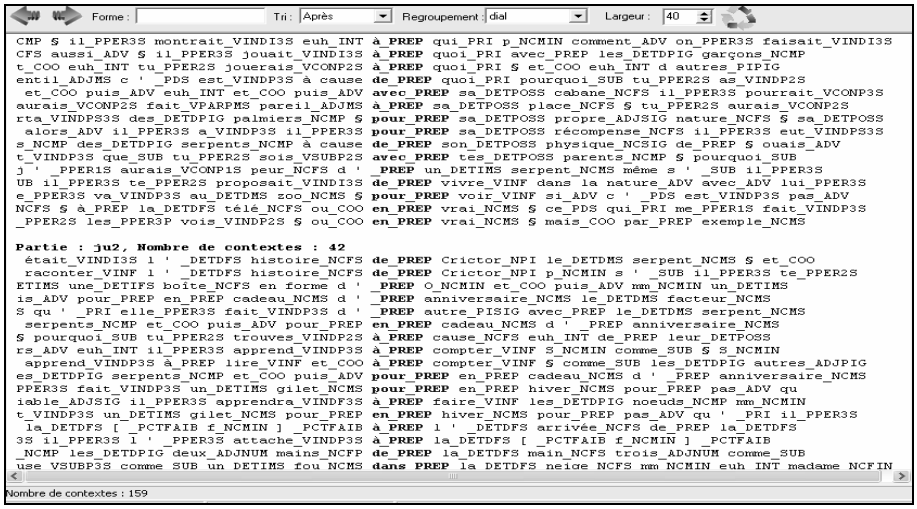
Pour approfondir les analyses, nous passons les corpus de Julien et de Mathilde à l'étiqueteur Cordial. Nous utilisons ensuite un programme de préparation textuelle (MkCorpus)[6] pour extraire du corpus étiqueté, inutilisable pour Lexico3, deux corpus distincts au format approprié. Nous créons ainsi les corpus :

- JULIEN-forme-cat (§<ad=101> alors_ADV c'_PDS est_VINDP3S quoi_PRI l'_DETDFS histoire_NCFS de_PREP Crictor_NPI)
- JULIEN-lemme-cat (§<ad=101> alors_ADV ce_PDS être_VINDP3S quoi_PRI le_DETDFS histoire_NCFS de_PREP Crictor_NPI).

À partir de JULIEN-forme-cat, nous pouvons rechercher l'emploi des prépositions. Nous constituons un groupe de forme grâce à l'expression régulière « [a-zà]*_PREP », qui se traduit par « tous les mots commençant par une lettre dans l'intervalle a-z ou à, se répétant ou non, suivi de la catégorie Préposition ». Nous obtenons la liste suivante :



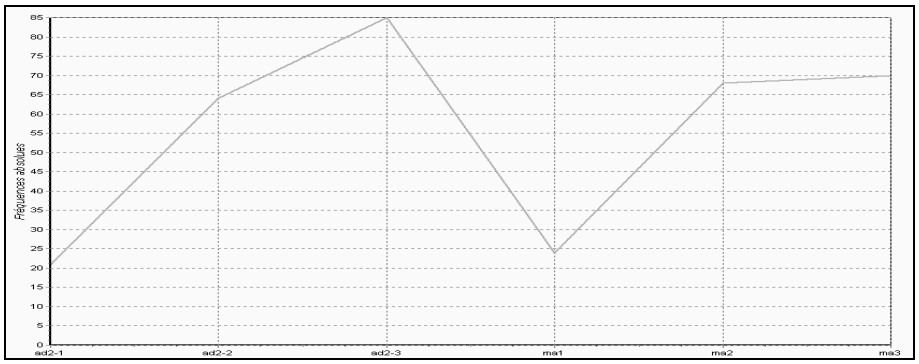
Nous procédons ensuite à la concordance de ce groupe pour vérifier le contexte immédiat des prépositions (Graphique 7). Nous pouvons ainsi vérifier les constructions de type « Verbe conjugué + Préposition », « Préposition + Verbe infinitif »… Nous pouvons trier les formes par ce qui précède ou ce qui suit, par dialogues ou par locuteurs.


Forme :
Tri : Après
Regroupement : dial
Largeur : 40
CMP $ il_PPER3S montrait_VINDI3S euh_INT à_PREP qui_PRI p_NCMIN comment_ADV on_PPER3S faisait_VINDI3S
CFS aussi_ADV $ il_PPER3S jouait_VINDI3S à_PREP quoi_PRI avec_PREP les_DETDPIG garçons_NCMP
t_COO euh_INT tu_PPER2S jouerais_VCONP2S à_PREP quoi_PRI $ et_COO euh_INT d autres_PIPIG
entil_ADJMS c ' _PDS est_VINDP3S à cause de_PREP quoi_PRI pourquoi_SUB tu_PPER2S as_VINDP2S
et_COO puis_ADV euh_INT et_COO puis_ADV avec_PREP sa_DETPOSS cabane_NCFS il_PPER3S pourrait_VCONP3S
aurais_VCONP2S fait_VPARPMS pareil_ADJMS à_PREP sa_DETPOSS place_NCFS $ tu_PPER2S aurais_VCONP2S
rta_VINDP3S des_DETDPIG palmiers_NCMP $ pour_PREP sa_DETPOSS propre_ADJSIG nature_NCFS $ sa_DETPOSS
alors_ADV il_PPER3S a_VINDP3S il_PPER3S pour_PREP sa_DETPOSS récompense_NCFS il_PPER3S eut_VINDPS3S
s_NCMP des_DETDPIG serpents_NCMP à cause de_PREP son_DETPOSS physique_NCSIG de_PREP $ ouais_ADV
t_VINDP3S que_SUB tu_PPER2S sois_VSUBP2S avec_PREP tes_DETPOSS parents_NCMP $ pourquoi_SUB
j ' _PPER1S aurais_VCONP1S peur_NCFS d ' _PREP un_DETIMS serpent_NCMS même s ' _SUB il_PPER3S
UB il_PPER3S te_PPER2S proposait_VINDI3S de_PREP vivre_VINF dans la nature_ADV avec_ADV lui_PPER3S
e_PPER3S va_VINDP3S au_DETDMS zoo_NCMS $ pour_PREP voir_VINF si_ADV c ' _PDS est_VINDP3S pas_ADV
NCFS $ à_PREP la_DETDFS télé_NCFS ou_COO en_PREP vrai_NCMS $ ce_PDS qui_PRI me_PPER1S fait_VINDP3S
_PPER2S les_PPER3P vois_VINDP2S $ ou_COO en_PREP vrai_NCMS $ mais_COO par_PREP exemple_NCMS
Partie : ju2, Nombre de contextes : 42
était_VINDI3S l ' _DETDFS histoire_NCFS de_PREP Crictor_NPI le_DETDMS serpent_NCMS $ et_COO
raconter_VINF l ' _DETDFS histoire_NCFS de_PREP Crictor_NPI p_NCMIN s ' _SUB il_PPER3S te_PPER2S
ETIMS une_DETIFS boîte_NCFS en forme d ' _PREP O_NCMIN et_COO puis_ADV mm_NCMIN un_DETIMS
is_ADV pour_PREP en_PREP cadeau_NCMS d ' _PREP anniversaire_NCMS le_DETDMS facteur_NCMS
S qu ' _PRI elle_PPER3S fait_VINDP3S d ' _PREP autre_PISIG avec_PREP le_DETDMS serpent_NCMS
serpents_NCMP et_COO puis_ADV pour_PREP en_PREP cadeau_NCMS d ' _PREP anniversaire_NCMS
$ pourquoi_SUB tu_PPER2S trouves_VINDP2S à_PREP cause_NCFS euh_INT de_PREP leur_DETPOSS
rs_ADV euh_INT il_PPER3S apprend_VINDP3S à_PREP compter_VINF S_NCMIN comme_SUB $ S_NCMIN
apprend_VINDP3S à_PREP lire_VINF et_COO à_PREP compter_VINF $ comme_SUB les_DETDPIG autres_ADJPIG
es_DETDPIG serpents_NCMP et_COO puis_ADV pour_PREP en_PREP cadeau_NCMS d ' _PREP anniversaire_NCMS
PPER3S fait_VINDP3S un_DETIMS gilet_NCMS pour_PREP en_PREP hiver_NCMS pour_PREP pas_ADV qu
iable_ADJSIG il_PPER3S apprendra_VINDF3S à_PREP faire_VINF les_DETDPIG noeuds_NCMP mm_NCMIN
t_VINDP3S un_DETIMS gilet_NCMS pour_PREP en_PREP hiver_NCMS pour_PREP pas_ADV qu ' _PRI il_PPER3S
la_DETDFS [_PCTFAIB f_NCMIN] _PCTFAIB à_PREP l ' _DETDFS arrivée_NCFS de_PREP la_DETDFS
3S il_PPER3S l ' _PPER3S attache_VINDP3S à_PREP la_DETDFS [_PCTFAIB f_NCMIN] _PCTFAIB
_NCMP les_DETDPIG deux_ADJNUM mains_NCFP de_PREP la_DETDFS main_NCFS trois_ADJNUM comme_SUB
use_VSUBP3S comme_SUB un_DETIMS fou_NCMS dans_PREP la_DETDFS neige_NCFS mm_NCMIN euh_INT madame_NCFIN
Nombre de contextes : 159

Graphique 7 : Concordance des Prépositions répartis par dialogues avec un tri par ce qui suit.

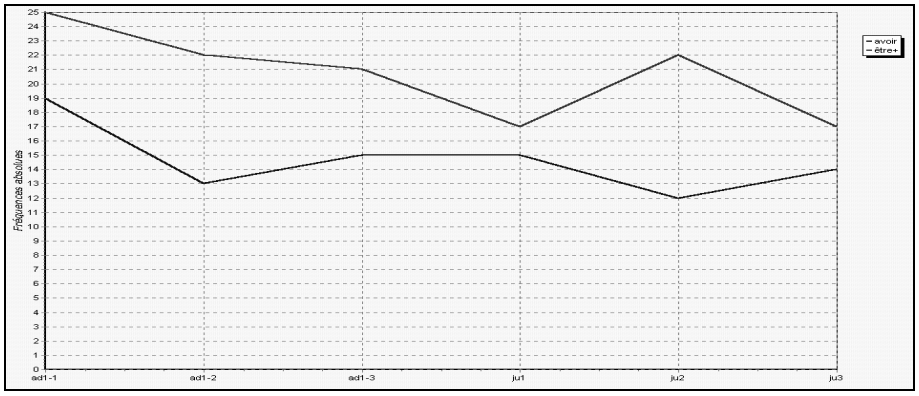
Avec MATHILDE-forme-cat, nous nous apercevons que l'emploi des prépositions est en progression constante (Graphique 8) et que cette progression se vérifie également pour chaque locuteur (Graphique 9).

Graphique 8 : Répartition par dialogues

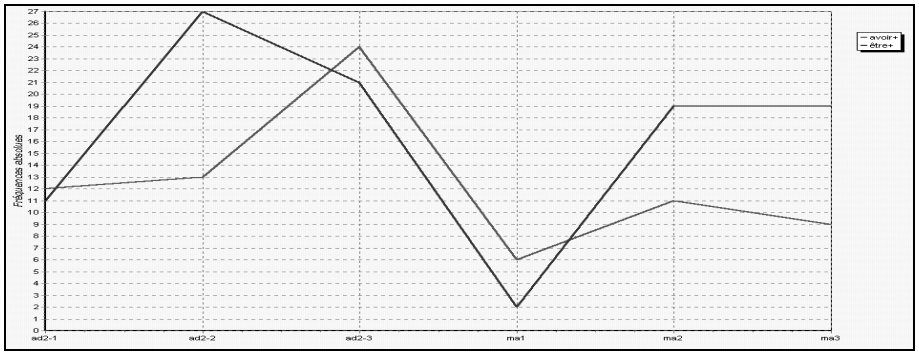


Graphique 9 : Répartition par locuteurs

À partir de JULIEN-lemme-cat, nous recherchons les emplois des verbes être et avoir. Nous remarquons que le verbe être est toujours plus utilisé que le verbe avoir (Graphique 10). La même recherche avec MATHILDE-lemme-cat nous montre que les fréquences sont croisées (Graphique 11).

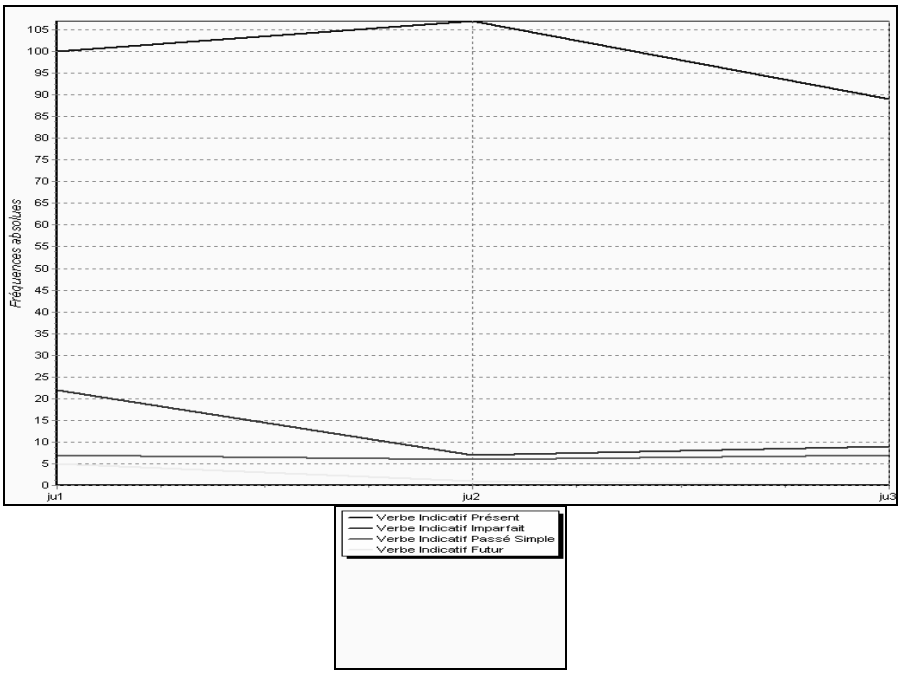


Graphique 10 : Répartition par locuteurs avec JULIEN-lemme-cat



Graphique 11 : Répartition par locuteurs avec MATHILDE-lemme-cat

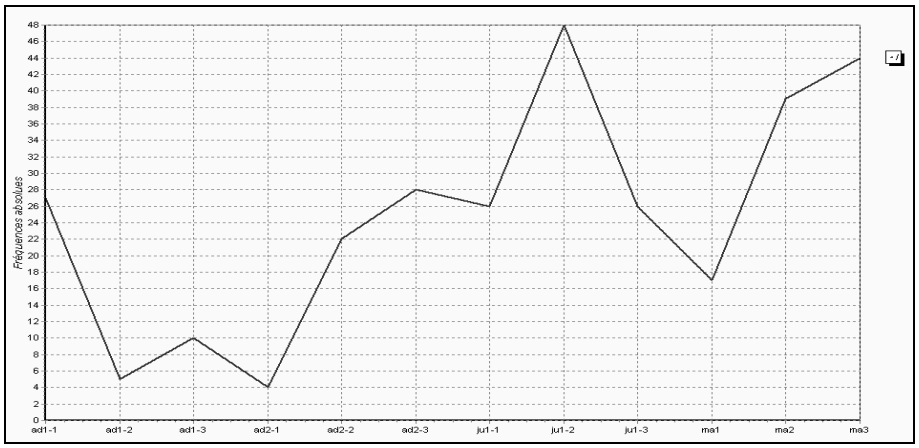
À l'aide des formes lemmatisées, nous pouvons rechercher le temps des verbes et comparer l'utilisation du présent par rapport à l'imparfait, comparer l'emploi des modes… (Graphique 12).



Graphique 12 : Répartition par dialogues avec JULIEN-lemme-cat

2.4. Corpus de Julien-Mathilde

Nous regroupons les corpus de Julien et de Mathilde en un seul corpus JulienMathilde. Cela permet de comparer directement les enfants entre eux, et l'adulte à lui-même. Nous pouvons alors rechercher, par exemple, les marques d'hésitations (/). Nous remarquons que l'adulte, dans le corpus de Mathilde (ad2) a de plus en plus de marque d'hésitation. Mathilde a également une croissance constante d'hésitation (Graphique 13).



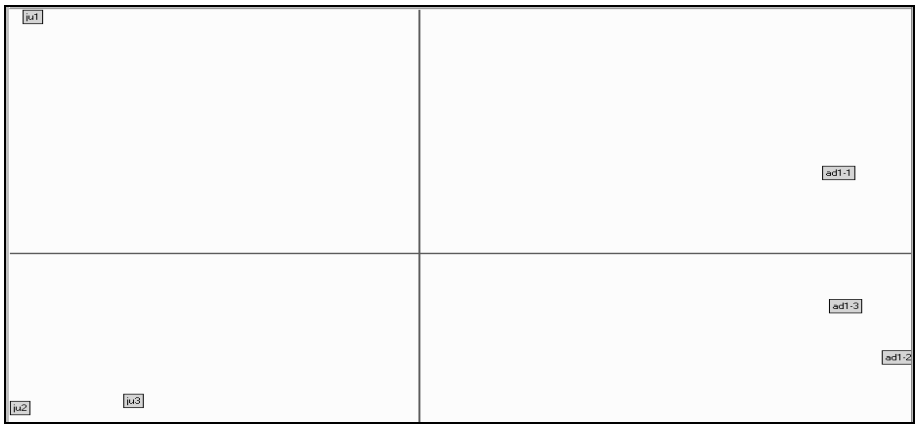
Graphique 13 : Répartition par locuteurs

Il faut ensuite revenir au texte par les concordances pour vérifier s'il s'agit d'une répétition de la forme, d'un tâtonnement sur une construction ou d'une tentative aboutie ou non d'une construction complexe.

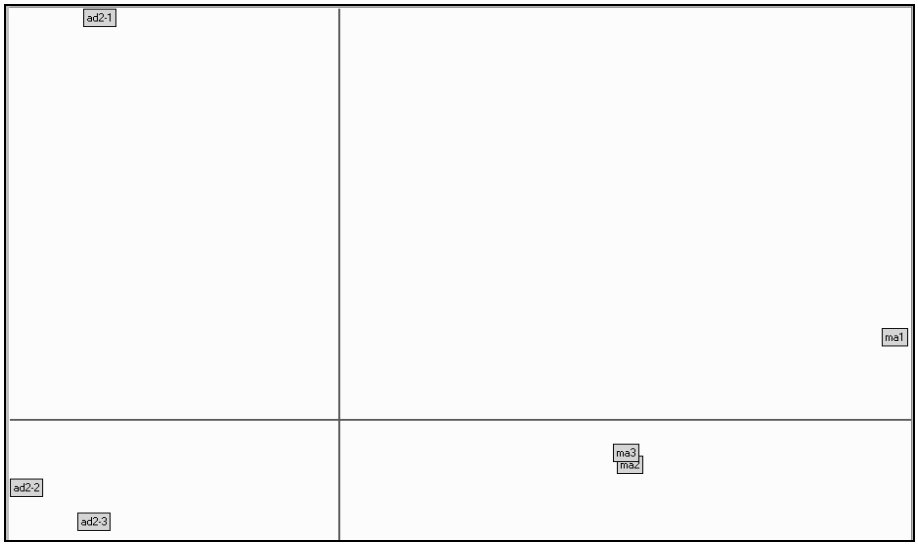
3. ANALYSES FACTORIELLES DES CORRESPONDANCES

Lorsque nous regroupons les corpus en un, nous pouvons comparer les enfants entre eux, ainsi que l'adulte à lui-même, et observer comment son discours peut évoluer en fonction de l'enfant à qui il s'adresse. Lexico3 permet de procéder à des analyses factorielles. L'analyse factorielle des correspondances fait partie de la famille d'analyse multidimensionnelle qui s'applique à des tableaux à double entrée. Cette méthode est caractérisée par le calcul de la distance « chi-2 » (Lebart et Salem, 1994).

Quand nous faisons l'analyse factorielle des correspondances (AFC) du corpus Julien en se basant sur la partition « locuteurs » (Graphique 14), nous remarquons deux groupes distincts (Julien à gauche et l'adulte à droite), mais un rapprochement des dialogues vers le centre. Nous observons le même phénomène de rapprochement avec les corpus étiquetés, ainsi que chez Mathilde (Graphique 15).

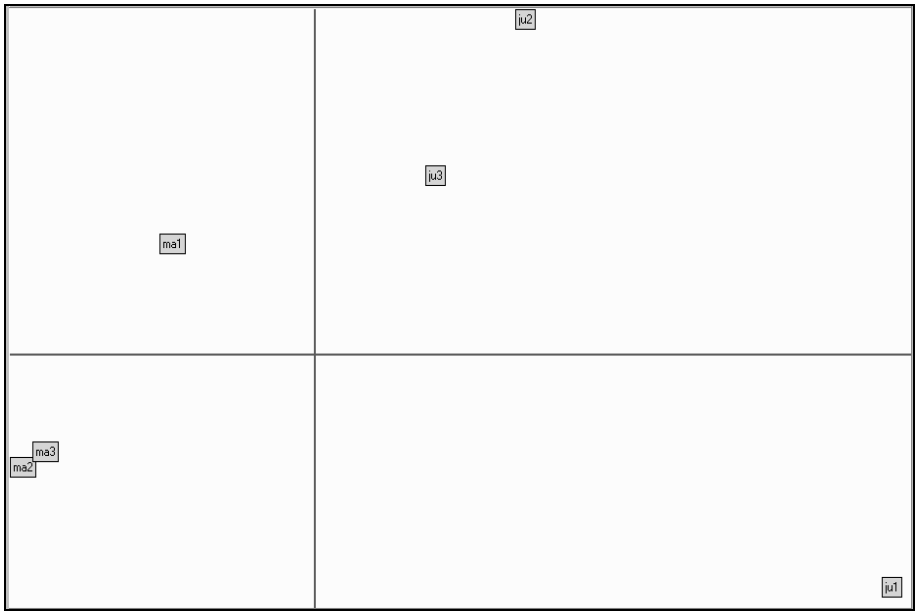


Graphique 14 : AFC de Julien par locuteurs

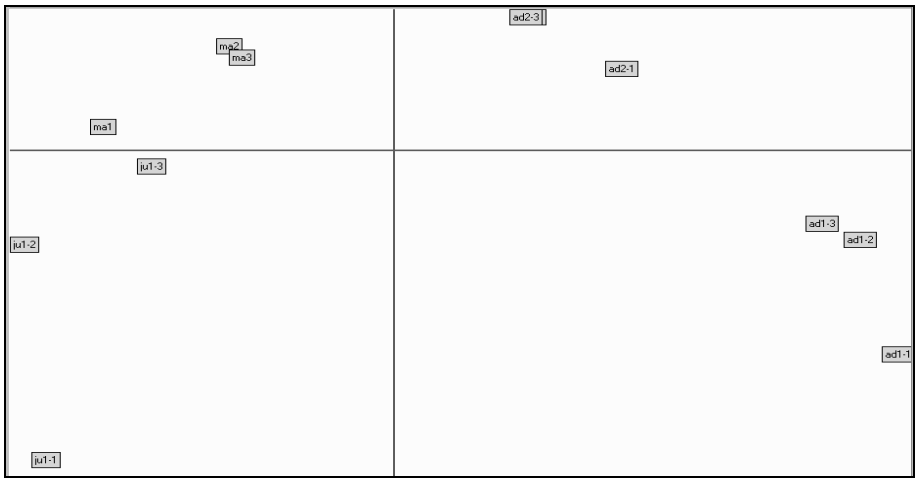


Graphique 15 : AFC de MATHILDE-forme-cat par locuteurs

Avec le corpus JulienMathilde, nous observons plusieurs choses. L'AFC par dialogues (Graphique 16) nous montre que le dernier dialogue de Julien est le plus proche du centre, et que le premier de Mathilde aussi. Mais il est à noter que dans la chronologie de Mathilde, il y a un rapprochement qui se fait dans le dernier dialogue. Si nous utilisons la partition « locuteurs » (Graphe 17), nous observons un rapprochement très distinct qui peut mettre en relief une adaptation du langage de l'adulte en fonction de l'enfant.



Graphique 16 : AFC de JulienMathilde par dialogues



Graphique 17 : AFC de JulienMathilde par locuteurs

Avec les AFC, il est possible d'observer un éloignement ou rapprochement de locuteurs. Nous avons montré qu'il est possible de mettre en relief qu'une adaptation du langage adressé à l'enfant le rapprochait du langage de l'adulte. « Il ne suffit pas à l'entourage d'exposer l'enfant à un modèle verbal défini une fois pour toutes : il s'agit de lui fournir le langage qui LUI convient, au stade particulier où il se trouve ; il importe aussi que ces données verbales s'intègrent dans un cadre où chacun communique avec

l'autre, ce qui implique, de la part de l'adulte, un intérêt pour l'enfant en tant que partenaire, une sensibilité à ses besoins, à ses intentions, à ses capacités, à ce qu'il comprend, à ce qu'il tente de communiquer » (Moreau et Richelle, 1981).

4. RÉFÉRENCES

Blanche-Benveniste C. 1997. *Approches de la langue parlée en français*. Paris : OPHRYS. Collection l'essentiel français.

Habert B, Nazarenko A. et Salem A. 1997. *Les linguistiques de corpus*. Paris : Colin.

Kerbrat-Orecchioni C. 1999. « L'oral dans l'interaction : une liberté surveillée », Revue Française de Linguistique Appliquée, Volume IV n°2, décembre 1999, (p. 41-55).

Lebart L. et Salem A. 1994. *Statistique textuelle*. Paris : Dunod.

Lentin L. 1998. *Apprendre à penser, parler, lire, écrire*. Paris : ESF.

Lentin L. *et al.* 1984. *Recherche sur l'Acquisition du Langage*, tome 1. Paris : Publications de la Sorbonne Nouvelle.

Lentin L. *et al.* 1988. *Recherche sur l'Acquisition du Langage*, tome 2. Paris : Publications de la Sorbonne Nouvelle.

Moreau M.-L. et Richelle M. 1981. *L'acquisition du langage*, Mardaga, (4ème édition).

Ringeard M. et Lorscheider U. 1977. « Edition d'un corpus de français parlé », Recherches sur le français parlé n°1, (p.15-24).

Sansonetti L. 2001. *Le Projet ANACAL*, Paris, Mémoire de Maîtrise, ILPGA, Paris 3, (non publié).

Sansonetti L. 2002. *Approche lexicométrique de corpus de dialogues adulte-enfant dans le cadre de recherche en linguistique de l'acquisition du langage*, Paris, Mémoire de DEA, ILPGA, Paris 3, (non publié).

Tissier C. 2001. *Rôle de l'adulte dans l'interaction langagière adulte-enfant (entre 4 ans 9 mois et 6 ans 4 mois) en situation de narrations dans deux corpus longitudinaux*, Paris, Mémoire de Maîtrise, ILPGA, Paris 3, (non publié).

5. NOTES

(1) Ungerer, T., *Crictor*, 1958, l'École des Loisirs pour l'édition en langue française, 1980, réédition 200 en collection « lutin poche ».
(2) Groupe Aixois de Recherches en Syntaxe.
(3) Centre de Recherche sur l'Acquisition du Langage Oral et Ecrit, rattaché au CALIPSO (EA 170 – Paris III).
(4) http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/
(5) http://synapse.com
(6) http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/mkcorpusProject.htm

LINGUISTIQUE ET CORPUS

NÉOLOGIE DU PORTUGAIS BRÉSILIEN

Ieda Maria Alves
Université de São Paulo

1. INTRODUCTION

La linguistique de corpus est devenue un outil irremplaçable pour l'exploitation de textes informatisés à plusieurs buts. Cette discipline, qui avance et prend sa place parmi les disciplines à caractère linguistique, est de plus en plus employée dans de différentes applications, telles que la lexicographie, la sociolinguistique, l´enseignement de langues, la traduction, la linguistique historique, la rédaction technique, la reconnaissance de la voix, la linguistique computationnelle, la morphologie dérivationnelle (McEnery *et al.* 2000 ; Battaner 2000 ; Plénat *et al.* 2002 ; Booker et Pearson 2002 ; Sardinha 2004).

Parmi les possibles applications de la linguistique de corpus, nous présentons quelques résultats auxquels nous sommes arrivés en explorant un corpus concernant un projet de recherche universitaire dénommé *Base de Néologismes du Portugais Brésilien Contemporain*. Ce projet, qui reçoit l'appui financier du Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), est intégré aussi par des étudiants de Maîtrise et de Doctorat.

La *Base de Néologismes du Portugais Brésilien Contemporain* a l'objectif général de fournir des éléments pour l'étude du lexique portugais (variante brésilienne) ainsi que pour l'élaboration de répertoires d'unités lexicales néologiques.

À part ces objectifs généraux, la *Base* a également pour objectif d'étudier :

- les procédés de formation les plus usuels dans la période et dans le corpus étudiés ;
- les formants affixaux (préfixes et suffixes) les plus productifs dans la période et dans le corpus étudiés ;
- la concurrence entre emprunts et éléments vernaculaires dans l'évolution du lexique portugais ;
- la circulation des termes des sciences et des techniques dans la presse brésilienne.

Ce projet de veille néologique s'appuie sur un corpus informatisé recueilli sur Internet et concernant la presse écrite brésilienne (journaux *Folha de S. Paulo* – FSP – et *O Globo* – G – et revues *Veja* – V – et *IstoÉ* – IE qui sont les plus diffusés dans le territoire brésilien). Ce choix n'est pas fortuit. Les revues *IstoÉ* et *Veja*, publiées à São Paulo et les plus diffusées au Brésil, traitent de divers domaines des langues de spécialité. Elles correspondent à un corpus de vulgarisation où des questions concernant plusieurs domaines de spécialité sont diffusées pour un public non-spécialisé. Les journaux *Folha de S. Paulo* et *O Globo*, publiés à São Paulo et Rio de Janeiro, respectivement, présentent aussi divers domaines (politique, sports, arts...) et sont également les plus diffusés au Brésil.

Ces matériaux sont analysés par échantillonnage et chacun est analysé une fois par mois :

1. journal *O Globo* - premier dimanche du mois ;
2. revue *IstoÉ* - deuxième semaine du mois ;
3. journal *Folha de S. Paulo* - troisième dimanche du mois ;
4. revue *Veja* - quatrième semaine du mois.

Dans ces périodiques, nous collectons, depuis janvier 1993, des néologismes de caractère vernaculaire (dérivation, composition, troncation, transfert sémantique...) et de caractère étranger. Parmi les critères souvent mentionnés pour la reconnaissance du statut néologique ou non d'une unité lexicale – instabilité formelle, perception de la nouveauté par les usagers, emploi récent, enregistrement lexicographique (Cabré, 1993) –, nous choisissons comme critère principal le filtre lexicographique, qui consiste à vérifier l'existence ou à constater l'absence de l'unité lexicale dans une série d'ouvrages lexicographiques. D'après ce critère, dénommé corpus d'exclusion (Boulanger, 1978), nous considérons néologiques les unités lexicales qui ne sont pas incluses dans les dictionnaires de langue :

FERREIRA, A. B. de H. (2 ed.) 1986. *Novo dicionário da língua portuguesa*. Rio de Janeiro : Nova Fronteira. (1 ed.) 1975 (pour le corpus correspondant à la période comprise entre 1993 et 1999).
___. (3 ed.) 1999. *Novo Aurélio Século XXI : o dicionário da língua portuguesa*, Rio de Janeiro : Nova Fronteira. (1 ed.) 1975 (pour le corpus recueilli à partir de 2000).
___.2001. *Dicionário houaiss da língua portuguesa*. Rio de Janeiro : Objetiva (pour le corpus recueilli à partir de 2002).
MICHAELIS. 1998. *Moderno dicionário da língua portuguesa*. São Paulo : Melhoramentos (pour le corpus recueilli à partir de 1999).

Le corpus d'inclusion comprend aussi le vocabulaire orthographique publié par l'Academia Brasileira de Letras, qui présente, dans sa

macrostrucure, un grand nombre d'unités lexicales qui ne sont pas intégrées à des dictionnaires de langue générale :

ACADEMIA BRASILEIRA DE LETRAS. 1981. *Vocabulário ortográfico da língua portuguesa*. Rio de Janeiro : Bloch (pour le corpus correspondant à la période comprise entre 1993 et 1997).
ACADEMIA BRASILEIRA DE LETRAS. (2 ed.) 1998. *Vocabulário ortográfico da língua portuguesa*. Rio de Janeiro : Academia. (1ère ed.) 1981 (pour le corpus recueilli à partir de 1999).

Cet ensemble de corpus de presse est dépouillé à l'aide du logiciel Folio Builder 4.2, logiciel américain qui nous permet d'effectuer des recherches par préfixes, par suffixes, par des radicaux, ainsi que d'établir des listes de fréquences et de concordances. Les unités lexicales inventoriées sont transcrites dans une fiche lexicale (à l'aide du logiciel Access de Microsoft) qui présente les champs suivants : unité lexicale ; sigle ou acronyme ; références grammaticales ; contexte(s) ; références du(des) contexte(s) ; domaine ; sous-domaine ; notes linguistiques ; notes complémentaires ; synonymes ; variantes ; auteur de la transcription ; date de la transcription. Les champs obligatoirement remplis sont les suivants : unité lexicale ; références grammaticales ; contexte(s) ; références du (des) contexte(s) ; notes linguistiques ; notes complémentaires ; auteur de la transcription ; date de la transcription.

2. QUELQUES RÉSULTATS

La *Base* présente, actuellement, 11 900 unités lexicales néologiques (résultats concernant la période comprise entre 1993 et 2000). Ces unités lexicales présentent une, deux ou plusieurs occurrences. On constate, donc, que quelques unités lexicales constituent des hapax, c'est-à-dire, des unités qui ne présentent qu'une seule occurrence, tandis que d'autres se montrent très fréquentes, en présentant plusieurs occurrences le long de la période étudiée, ce qui montre qu'elles sont en train de s'intégrer à la langue portugaise. Le nombre d'occurrences de la *Base* représente près de 19 300 contextes.

Les données déjà collectées nous permettent d'observer quelques faits concernant les aspects morphologiques des unités lexicales néologiques. Nous pouvons déjà conclure à propos de la productivité des procédés de formation des matériels dépouillés, qui se présente d'après le schéma suivant (cf. tableau 1) :

- dérivation par préfixation (30% des données) ;
- composition par subordination (19% des données) ;
- emprunt (17% des données) ;

- formation syntagmatique (13% des données) ;
- dérivation suffixale (8% des données) ;
- composition par coordination (5% des données) ;
- néologie sémantique (3% des données) ;
- d'autres procédés (5% des données).

Ces résultats montrent que, dans le corpus analysé, les néologismes vernaculaires – les dérivés par préfixation et les composés par subordination – sont les procédés les plus employés et correspondent à 83% des unités lexicales néologiques. Les mots étrangers ne correspondent qu'à 17% des unités lexicales néologiques, dont 68% constituent des hapax.

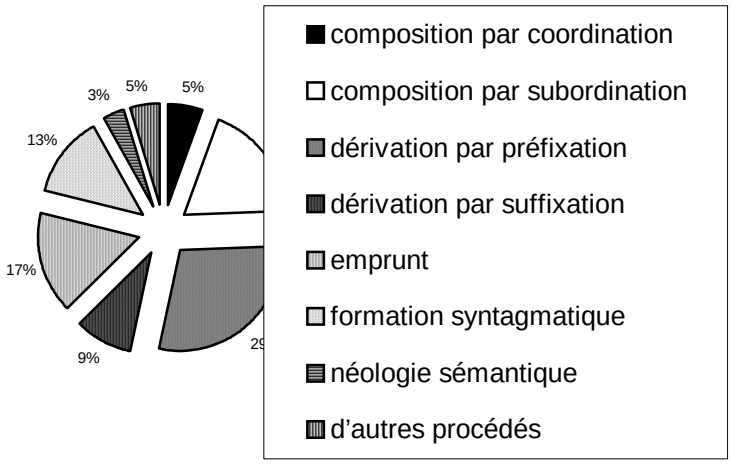


Tableau 1 : Procédés d'innovation lexicale

D'autres résultats concernent l´emploi des formants affixaux (préfixes et suffixes) les plus productifs dans la période et dans le corpus étudié.

L'analyse des dérivés suffixaux nous démontre que les suffixes les plus productifs pour la formation de nouvelles unités lexicales sont *-ismo*, *-ista*, *-ção* et *-ano* (tableau 2) :

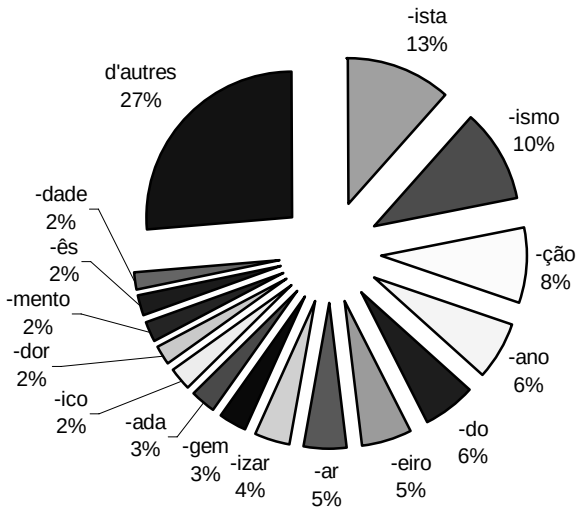


Tableau 2 : Suffixes

Les préfixes les plus productifs concernent la négation et l'opposition (*anti-*, *não-*) et l'intensité (*super-*) :

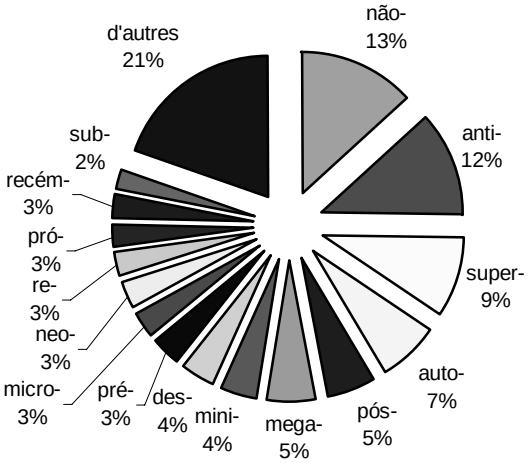


Tableau 3 : Préfixes

Les unités lexicales étrangères les plus employées dans ce corpus de presse analysé sont d'origine anglaise (79%), concernant surtout la musique, les sports, l'économie et les technologies. Les autres langues dont le portugais brésilien emprunte des mots sont le français (6%), suivi du japonais (3%), de l'espagnol (2%) et de l'italien (2%).

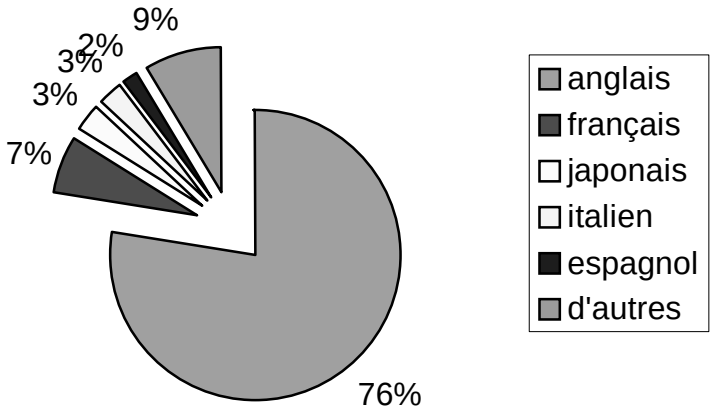


Tableau 4 : Xénismes

Un autre résultat concerne les langues de spécialité. Les unités lexicales du domaine de l'économie sont les plus employées (31%) dans le corpus étudié. Aussi fréquentes sont les unités lexicales qui représentent la musique, la gastronomie et les technologies (tableau 5).

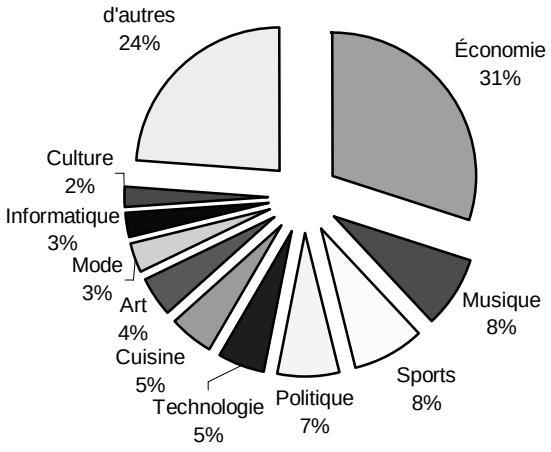


Tableau 5 : Domaines

Parallèlement à ces résultats quantitatifs, les données de la *Base* permettent déjà l'étude de plusieurs recherches, qui concernent surtout des aspects morphologiques, tels que la concurrence entre les suffixes *-ista* et *-eiro* dans la formation d'unités lexicales désignatives de profession, de *-mento* et *-ção* (Maroneze et Nascimento 2001 ; Maroneze 2002), la composition avec des formants radicaux grecs et latins (Oliveira et Rosiska, 2002), le rôle de la métaphore dans la formation de néologismes sémantiques (Alves, Rosiska et Maroneze, 2002), la concurrence entre emprunts et formations vernaculaires (Alves et Maroneze, 2002). En voici des exemples, qui démontrent la concurrence entre les suffixes *-mento* et *-ção :*

> *Quem se submete às cirurgias não está cometendo nenhum crime, pois não há punição para a <autoflagelação>. Mas ocorre que muitos transexuais acabam conseguindo documentação falsa.* (FSP, 20-03-94)[1]
>
> *A sociedade brasileira sempre fez uma leitura negativa de si, de <autoflagelamento>. Como se nos Estados Unidos e na Alemanha não houvesse hipocrisia com o trabalho.* (IE, 13-11-96)[2]

Les données de la *Base* montrent aussi quelques faits qui contrarient, d'une certaine façon, quelques aspects historiques de la langue portugaise. Dans l'histoire de cette langue, les formations par suffixation sont très nombreuses par rapport aux autres modes de formation, étant même considérées comme d'extrême importance pour la formation d'un lexique

typiquement portugais, du XIII^{e} au XIV^{e} siècles (Coelho et Silva, 2004) ; néanmoins, les résultats auxquels nous sommes arrivés montrent que, synchroniquement, les préfixes sont beaucoup plus employés que les suffixes dans la formation de nouvelles unités lexicales (cf. tableau 1).

Les données de la *Base* révèlent aussi que la grande quantité de formations par dérivation préfixale est la conséquence du fait que de nombreux formants d´origine grecque et latine (*auto*-, *hiper*-, *mega*-, *macro*-, *micro*-, *multi*, *neo*-...), qui auparavant n´étaient employés que dans le langage scientifique, sont maintenant employés plutôt dans la langue générale, en s´associant à des formants radicaux qui n´appartiennent pas au langage scientifique. Ces formants d´origine grecque et latine, auparavant classés comme des composés, sont plutôt considérés comme des préfixes – classification suivie dans ce travail – car ils circulent dans la langue générale. Un exemple de ce fait peut être illustré par le formant *hiper*-, qui est enregistré dans les dictionnaires de langue comme formateur de termes de plusieurs terminologies (médecine, géométrie, physique...) et qui, dans le corpus analysé, forme notamment des unités lexicales qui n´appartiennent pas à un domaine spécialisé. En voici quelques exemples :

> *A revista "Time" prevê, entre outras coisas, a chegada do homem a Marte em 2017 e a construção nos próximos 20 anos de um <hiperavião> que levará mil passageiros a uma distância de 10.000 Km a 900 Km/h.* (G, 06-set-98) [3]

Les nombreuses formations par préfixation sont explicables aussi par certains préfixes qui révèlent des soucis contemporains (l´écologie, par exemple, qui nous fournit le formant préfixal *eco*-) ou le développement de l'informatique, dont les formants *ciber*- et *e*- sont des exemples :

> *O primeiro espaço <eco-cultural> do Rio vai ser inaugurado na Tijuca, na esquina da Garibaldi com Conde de Bonfim, com projetos de meio ambiente.* (G, 02-05-99)[4]
>
> *<Ciberguerra> (tít.)*
>
> *Novo lance na luta pelo domínio dos programas de navegação na Net.* (IE, 14-05-97)[5]
>
> *O novo gênero, que inclui outros sucessos como "Matrix" /.../ está sendo chamado <"e-cinema"> (cinema eletrônico ou digital).* (FSP, 20-jun-99)[6]

On remarque aussi que, dans les syntagmes nominaux, une concurrence s´installe entre le déterminant adjectival et le déterminant à caractère prépositionnel (*mercado de câmbio* / *mercado cambial*) :

> *O que também beneficia o <mercado de câmbio>, são as emissões de títulos brasileiros no exterior.* (G, 06-08-00)[7]
>
> *No <mercado cambial>, o dólar comercial encerrou os negócios na sexta com nova cotação recorde, vendido a R$ 2,245.* (FSP, 22-04-01)[8]

Comme conséquence de cette concurrence, les exemples recueillis semblent montrer une préférence pour l´emploi de l´adjectif, au détriment de la préposition. Cette perte des prépositions est aussi remarquée dans la formation de composés formés par deux noms, tels que *salário-maternidade*, où l´on remarque un effacement de la préposition *de* :

> *Parlamentares mulheres foram ao STF pedir o fim do limite para o <salário-maternidade>*. (V, 31-03-99) [9]

3. CONSIDÉRATIONS FINALES

Nous avons essayé de montrer, dans cet exposé, que l´exploitation de corpus pour la description des aspects morphologiques d´une langue est aussi une des applications possibles de la linguistique de corpus. Les exemples présentés ne sont, en fait, qu´un petit échantillon de tous les apports que cette discipline peut procurer à la description morphologique d´une langue.

4. RÉFÉRENCES

Alves I. M. ; Maroneze B. O. et Rosiska P. 2002. « Metáfora e criação lexical num corpus jornalístico ». *Conference on metaphor in language and thought*. São Paulo : PUC-SP. (p. 24).

Battaner M. P. 2000. « Un corpus para la ensenanza : Corpus PAAU, junio, 1992 ». *VI Jornada de corpus lingüístics*. Barcelona : IULA-Universitat Pompeu Fabra. (p. 41-64).

Boulanger J.-C. 1979. « Néologie et terminologie ». *Néologie en Marche*. N° 4. (p. 9-127).

Bowler L. et Pearson J. 2002. *Working with specialised language*. London, New York : Routledge.

Cabré M. T. 2003. *La terminología. Teoría, metodología, aplicaciones*. Barcelona : Editorial Antártida, Empúyries.

Coelho J. S. B. et Silva R. V. M. 2004. « A sufixação e a formação do léxico português arcaico ». ANPOLL Boletim Informativo n° 32. (p. 234).

Maroneze B. O. et Nascimento V. C. C. 2001. « A neologia do português contemporâneo do Brasil – a derivação sufixal ». 9 Simpósio Internacional de Iniciação Científica. Université de São Paulo.

Maroneze B. O. 2002. « Neologismos formados por sufixos indicadores de ação ». 10 Simpósio Internacional de Iniciação Científica. Université de São Paulo.

McEnery T., Wilson A., Baker P. 2000. « Linguistic corpora and language teaching : corpus-based help for teaching grammar ». VI Jornada de corpus lingüístics. Barcelona. IULA-Universitat Pompeu Fabra. (p. 65-76).

Plénat M., Lignon S., Serna N., Tanguy L. 2002. « La conjecture de Pichon ». Meillet, S. (dir.). Corpus et recherches linguistiques. Nice : Publications de la Faculté des Lettres, Arts et Sciences Humaines de Nice. (p. 105-50).

Sardinha T. B. 2004. *Lingüística de corpus*. São Paulo : Manole.

5. NOTES

(1) Celui qui subit de telles chirurgies ne commet aucun crime, car il n´y a pas de punition pour l´<autoflagelação>. Il arrive, néanmoins, que de nombreux transsexuels arrivent à obtenir de faux documents.

(2) La société brésilienne a toujours fait une lecture négative de soi-même, de l´<autoflagelamento>. Comme si aux États-Unis il n´y avait pas d´hypocrisie par rapport au travail.

(3) La revue *Time* prévoit, parmi d'autres événements, l'arrivée de l'homme sur Mars en 2017 et la construction dans les prochains 20 ans d´un <hiperavião> qui transportera mille passagers dans un parcours de 10.000 Km à 900 Km/h.

(4) Le premier espace <eco-cultural> de Rio sera inauguré à l'angle de la rue Garibaldi avec la rue Conde de Bonfim, à Tijuca, présentant des projets de l'environnement.

(5) <Ciberguerra> (titre)
Nouvelle étape dans la lutte pour le domaine des logiciels de navigation sur le Net.

(6) Le nouveau genre, qui inclut d´autres réussites telles que « Matrix » /.../ est à présent appelé < « e-cinema » > (cinéma électronique ou numérique).

(7) Ce qui bénéficie également au <mercado de câmbio> sont les émissions de titres brésiliens à l´étranger.

(8) Dans le <mercado cambial>, le dollar commercial a atteint vendredi une nouvelle cotation record, vendu à R$ 2,245.

(9) Des femmes du Parlement sont allées au STF pour demander la fin de la limite pour le <salário-maternidade>.

OBSERVATIONS LINGUISTIQUES SUR UN CORPUS DE LÉGENDES EN ANGLAIS DE PHOTOGRAPHIES DE TRAINS

Pierre J. L. Arnaud, François Maniez
Université Lumière-Lyon II, CRTT

1. INTRODUCTION

La recherche présentée ici est à l'intersection de la linguistique de corpus et de l'analyse de genre. Le genre étudié est constitué par les légendes de photographies de trains publiées dans des recueils et des magazines pour amateurs de chemins de fer. Un genre est une catégorie de textes présentant des caractéristiques référentielles (ici, la description du contenu de photographies de scènes ferroviaires) et des caractéristiques linguistiques (terminologie spécifique, tendances grammaticales comme la présence du passif dans les articles expérimentaux). Le genre correspond en outre à une communauté de discours, et les amateurs de chemins de fer sont les consommateurs des ouvrages et magazines qui constituent le support de celui qui nous intéresse. S'ils sont britanniques, ils sont souvent d'anciens *train-spotters*, dont ils ont le caractère obsessionnel et le souci du détail, et sont souvent fort savants, ce en quoi ils ne se distinguent d'ailleurs pas d'autres communautés, comme celles des amateurs de jazz ou de cyclisme. Ceci entraîne une exigence de technicité et de précision de la part des éditeurs ; nous ne sommes cependant en aucun cas dans le cadre d'une littérature technique qui serait destinée aux professionnels des chemins de fer, et qui ne contiendrait pas de photographies du type de celles dont nous étudions les légendes.

Étant donné la brièveté des textes et le caractère fortement contraint du contenu référentiel, nous proposons le terme *nano-genre* pour l'objet de notre étude. Parmi les nano-genres déjà analysés, on peut citer les annonces matrimoniales (Stubbs, 2001), les lettres de rédacteurs en chef de périodiques scientifiques aux auteurs impétrants pour leur indiquer le résultat de leur soumission (Flowerdew & Dudley-Evans, 2002), ou encore les remerciements en tête d'ouvrages ou articles scientifiques, dont Giannoni (2002) montre que dans certaines disciplines ils servent en fait surtout à

l'auto-promotion des auteurs, ce qui ne sauterait pas nécessairement aux yeux du non-initié. Notre nano-genre est loin des productions du monde universitaire, très étudiées depuis Swales (1990) en partie pour des raisons sociologiques, les chercheurs étant souvent des linguistes en poste dans des établissements de sciences "dures" ou technologiques. En particulier, on y chercherait en vain des actes de langage à visée sociale sauf dans d'assez rares cas de l'humour pouvant établir une connivence avec le lecteur.

Les publications dépouillées (réparties typologiquement entre ouvrages et magazines et dialectalement entre Grande-Bretagne, Australie et Nouvelle-Zélande d'une part et États-Unis d'autre part afin d'équilibrer le corpus) ont en commun de comporter de nombreuses photographies occupant une proportion importante des pages. Dans certains cas, le seul texte est celui des légendes ; dans d'autres cas, certaines sections comportent du texte non directement lié aux photographies, celles-ci accompagnées de leurs légendes constituant les autres sections. Un fil conducteur organise généralement le choix et la disposition des photographies, comme, dans l'un des ouvrages, la vie d'un photographe, ou dans un autre l'histoire ferroviaire d'une région. Notre nano-genre partage avec les remerciements cités ci-dessus la caractéristique de ne pas être un genre autonome ; en effet, si on peut imaginer des photographies sans légendes, l'inverse est évidemment impossible.

Pour restreindre la terminologie rencontrée et resserrer l'éventail des contenus, nous nous sommes cantonnés aux photographies de trains à vapeur. La photographie prototypique montre une locomotive prise de 3/4 avant, tractant un train en rase campagne, mais de nombreuses variations sont présentes, présentant un panorama complet de l'époque de la traction vapeur : manœuvres, train à l'arrêt en gare, etc. Nous utiliserons le terme *événement* pour référer à ce qui est saisi par la photographie.

Quelques considérations sur les photographies s'imposent ici, du fait du caractère secondaire du nano-genre par rapport à elles. Une photographie d'un train en marche prélève une phase du mouvement et la fige en une représentation en deux dimensions, immobile, silencieuse et inodore, souvent en noir et blanc. L'amateur reconstitue la scène d'une part parce qu'il sait lire les images, mais aussi parce que ses connaissances du sujet lui permettent d'appréhender certains indices spécifiques : par exemple, le nuage de fumée sortant de la cheminée est un indicateur du travail de la machine et donc du bruit produit, et sa forme dépend de la vitesse. Au besoin, le contenu de la légende viendra attirer l'attention sur tel ou tel détail ou apportera une information que la photographie est impuissante à préserver, comme le tintement d'une cloche ou le froid qui régnait.

Certaines légendes sont limitées à une phrase, comme l'exemple suivant :

> (1) Rebuilt "Patriot" 4-6-0 *Illustrious* heads a down express near Lancaster in 1962.

D'autres sont plus longues et comportent plusieurs phrases, comme la suivante, de longueur moyenne, extraite du même ouvrage :

> (2) Fowler produced some of the LMS Class 3 2-6-2Ts with condensing apparatus in an attempt to reduce exhaust emissions in long tunnel sections such as the London Metropolitan "widened lines". No 40024 was pictured at Moorgate between duties in 1959, with early LT stock as a background.

Dans ce cas, il y a presque toujours une phrase, que nous dénommerons *phrase-noyau*, qui décrit à elle seule l'événement. Celle de l'exemple (2) est soulignée. Les autres phrases apportent des informations extérieures à l'événement, comme des détails sur la biographie de la locomotive, sur sa classe, sur les pratiques ferroviaires, sur le parcours d'un train, sur le site, etc. Les légendes peuvent être liées à une photographie exclusivement, ou bien peuvent comporter un renvoi à une photographie voisine, comme dans le (très bref) exemple suivant :

> (3) More wheat moves to tidewater.

où le phorique renvoie à une image précédente plus amplement légendée représentant elle aussi un train de blé. Les légendes de deux ou trois photographies peuvent encore former un texte continu, des indications spatiales du genre de *above* venant supprimer toute ambiguïté référentielle. Nous avons fractionné les légendes de ce type.

La numérisation a porté sur 525 phrases-noyaux, dont la longueur moyenne en nombre de caractères est de 113,52, avec un écart-type de 50,02 (max. 254, min. 12). Les phrases-noyaux ont été soumises à des traitements simples, recherches par Word et tris sur le SGBD Fox Pro.

Presque toutes comportent deux informations quasiment obligatoires, la date de l'événement, jour précis, mois ou année, et l'identité détaillée de la locomotive. Les exemples ci-dessus montrent des syntagmes nominaux référant aux machines tout à fait représentatifs. Comme les syntagmes concernés peuvent être très longs, sans apporter d'information intéressante pour nos analyses, nous remplaçons dans ce qui suit l'identité de la locomotive par *X* et la date par *D*.

Le corpus manifeste une grande variété dans la présentation de l'événement. La phrase-noyau peut décrire l'événement directement, comme en (4) :

(4) In a typical suburban working, X approaches Roma Street with a set of side-loading cars.

Ailleurs, il est explicitement indiqué qu'il s'agit d'une photographie, d'une scène ou d'une vue :

(5) In a view from D the X, by then owned by the RGS although its tender still bore D&RGW lettering, eases down Dallas Divide toward Ridgway with a train of sheep on their way to winter pasture.

Nous parlons alors de *mise en perspective*. Le *verbe d'événement* est la forme qui dans la phrase-noyau réfère à l'événement. Dans l'exemple suivant, du fait de la mise en perspective, le verbe d'événement (souligné) est à une forme non-conjuguée, cas minoritaire dans le corpus (concernant 10% des 511 phrases-noyaux contenant un verbe d'événement) :

(6) On D, this locomotive was captured approaching North Blyth with coal for shipment at the nearby staithes.

La faible étendue des textes et la forte contrainte de contenu rendent un tel nano-corpus apte à faire ressortir des paramètres ayant des conséquences linguistiques repérables. Parmi les nombreux points linguistiques dignes d'intérêt, nous discuterons de faits de genre grammatical, et, du côté du lexique, de l'expression de l'immobilité et du mouvement.

2. GENRE GRAMMATICAL

On sait que, si l'anglais n'a pas de genre grammatical à proprement parler et que les phénomènes d'accord du type "genre" ne s'y manifestent que dans les phoriques de troisième personne du singulier, il existe quelques manifestations d'indétermination. Ceci concerne d'une part les animés : enfants en bas âge dont le sexe n'est pas connu (reprise de noms tels que *child*) et animaux supérieurs dont le sexe est connu, ainsi que les pays considérés comme des personnes morales. Par ailleurs, la référence à des bateaux ou navires se fait normalement par des pronoms féminins. Quirk *et al.* (1985) attribuent ceci à l'existence d'une "attitude affectueuse" et semblent, dans le cas des navires, admettre la variation, illustrée par l'exemple suivant, qui sent fortement le fabriqué[1] : *That's a lovely ship ! What is she/it called ?* Biber *et al.* (1999) indiquent la possibilité de variation, mais, alors que leur ouvrage est riche en statistiques, il n'en donne pas sur ce point. En fait, la variation semble concerner, plutôt que les bateaux, massivement référés au féminin, certains noms de machines et véhicules, que malheureusement aucune des trois grosses grammaires britanniques récentes ne prend la peine d'énumérer ("*other kinds of inanimates, such as cars*" pour Huddleston et Pullum (2002), "*inanimate entities such as ships*" pour Quirk *et al.* (1985)). Les locomotives à vapeur,

objet de tout l'amour des Britanniques et plus encore des amateurs de chemins de fer, sont en principe concernées au premier chef.

Soixante-neuf phrases-noyaux de notre corpus comportent un ou plusieurs anaphoriques renvoyant à un SN référant à une locomotive. Exemples :

> (7) X from Burton looked quite presentable as it shot past Lenton South Junction with an unidentified Class B passenger train from Nottingham.
> (8) X is on the home stretch toward the depot as she steams into town with the *Dixie Flyer* circa D.
>
> La répartition est la suivante :
>
> it/its : 62
> she/her : 7

Nous avons donc bien là une indétermination grammaticale à l'intérieur du nano-genre : il est impossible de prédire autrement que sous forme de probabilité quel anaphorique renverra au SN référant à une locomotive. On constate aussi que, dans un environnement où l'affection vis-à-vis de l'objet central des ouvrages et des photographies aurait pu se manifester par l'emploi d'anaphoriques féminins, ce cas est très minoritaire. Par ailleurs, les sept féminins se trouvent dans trois des sources, toutes américaines. On peut voir là une manifestation du plus grand recours au lexique familier ou au pittoresque des légendes des recueils américains de photos de trains. C'est en effet presque exclusivement dans ceux-ci qu'on trouve des traces d'humour et de familiarité, telle cette classe de locomotives trapues dénommées *fatsoes* par l'auteur d'une légende ! Lorsqu'on examine les 16 phrases-noyaux manifestant ouvertement de l'animisme, c'est-à-dire prêtant à des locomotives des caractéristiques d'animés, et dont 13 sont justement américaines, on constate que, sur les quatre comportant un anaphorique, deux ont un féminin (donc 1 sur 2 contre 1 sur 8,86 pour l'ensemble[2]). Tout autant que l'"attitude affectueuse" de Quirk *et al.* (1985), l'animisme, fait de style, présent dans le corpus constitue une explication des féminins. Certes, pour le francophone lisant l'anglais, les noms français *locomotive* et *machine* étant grammaticalement féminins, il n'y a pas de choc sémantique majeur à ce qu'une locomotive soit l'objet d'une anaphore par un féminin. Mais de quel sexe sont les locomotives pour les anglophones ? Les anaphoriques y référant peuvent être féminins comme on vient de le voir, et nous n'avons aucun cas de masculin dans le corpus. En outre, l'anglais n'ayant pas de lexème de langue générale dénotant "frère-ou-sœur"[3], on est obligé de choisir, et c'est *sister* qui désigne des machines du même type :

> (9) A sister 2-8-2, X, shoves hard at the rear.

(3 occurrences de *sister*, uniquement dans cet emploi, aucune de *brother*). Ceci est-il spontané, ou bien contraint par la préexistence des reprises de *ship* au féminin et du lexème *sister ship* ? En fait, *sister* est employé phoriquement plus largement que pour des antécédents reprenables au féminin, ce dont le sens 5 du *Collins-COBUILD English Dictionary* rend parfaitement compte[5]: "*You can use **sister** to describe something that is of the same type or is connected in some way to another thing you have mentioned*" et ceci est confirmé par l'exemple suivant extrait du texte suivi d'un des magazines-sources :

(10) The bridge was similar to its five sisters.

Une recherche sur *Google* montre que vis-à-vis de 9 470 occurrences de *its brother*, il y en a 198 000 de *its sister*, en grande partie explicables par cet emploi. Outre *city, ship*, ce sont en fait surtout des noms d'institutions ou de publications qui sont concernés.

La reprise par *sister* ne peut donc nous fournir une indication fiable sur le sexe "psychologique" des locomotives. Par ailleurs, le corpus comporte quand même un exemple d'animisme où une machine est personnifiée, et c'est comme homme (dans un ouvrage où l'on trouve à un autre endroit une reprise par *her*) !

(11) A trainload of fresh perishables is hurried toward eastern appetites behind an old man of the Espee. (*Espee = SP = Southern Pacific* – noter par ailleurs la métonymie qui ajoute au pittoresque).

3. LEXIQUE

3.1. Stationnement

Les verbes d'événement (voir plus haut) sont sans aucun doute la caractéristique lexicale la plus remarquable du nano-genre légende de photographies de train. Que se passe-t-il lorsque la locomotive de la photographie est à l'arrêt ? Une recherche onomasiologique dans le corpus nous donne les verbes suivants, dont toutes les occurrences ont la machine comme siège du procès :

pause (9 occurrences)
stand (5)
wait (3)
sit (2)

auxquels on peut rajouter *simmer* (1 occ.), verbe référant indirectement à l'immobilité via la chauffe de la chaudière, et *sun oneself* (1), touche d'animisme.

L'examen des photos ainsi décrites ne permet pas de déceler de différences entre les situations référées qui correspondraient aux différents verbes. Devant une situation identique, l'énonciateur a le choix entre diverses conceptualisations, ici comme une attente, une pause (donc une interruption momentanée), ou une position/attitude. Il y a donc une indétermination conceptuelle qui n'a d'ailleurs rien de bien remarquable. Plus remarquable est la concurrence entre *sit* et *stand*, car il s'agit de deux verbes d'attitude immobile, à rapprocher d'ailleurs de *lie*, non attesté dans le corpus, et on sait que l'anglais contraint bien plus que le français l'indication explicite du type de position. *Stand* et *sit* sont deux verbes dont l'analyse sémantique est délicate en raison de leur grande polysémie. Il est clair que les points de départ de ces polysémies sont la dénotation des attitudes humaines "être debout" et "être assis", et que, par métaphore, on en arrive à la dénotation de la position de référents inanimés incapables d'être "debout" ou "assis", ce qui rend impossible une traduction littérale vers le français (**La locomotive est debout dans la gare*). Le *Macmillan English Dictionary for Advanced Learners* (*MEDAL*) distingue pour *stand* un sens 5 "stationner"[4] : "*If a car, train, plane, etc. stands somewhere, it remains there without moving, waiting to be used*", alors que rien de similaire n'existe pour *sit*. Ce que font les dictionnaires, c'est opérer des coupes à des endroits saillants du continuum polysémique de tels verbes, et le sens 5 de *stand* pour le *MEDAL* résulte en fait de l'interaction du verbe avec un sujet référant à un véhicule. Pourquoi *stand* a-t-il pu prendre un tel sens et non *sit* ? C'est probablement encore par métaphore. En effet, on passe directement de la position debout à la marche ou à la course, alors qu'un homme assis doit d'abord se lever pour pouvoir se déplacer. *Stand* dénote donc une position contrastant directement avec le mouvement.

Seulement, comme on vient de le voir, le SN référant à une locomotive peut être le sujet de *sit*. Deux sens de *stand* et *sit* décrits par le *MEDAL* sont intéressants ici : *stand-* (4) *If an objects or building stands somewhere, it is in a particular position* (ex.: *Their house stood at the top of a hill*) ; *sit-* (3) *to be in a particular place* (ex. *The house sits on top of a hill overlooking the countryside*). À en juger par les définitions et les exemples, il n'y a ici aucune différence, et nous avons affaire à des synonymes dénotationnels, dès lors que les sèmes [DEBOUT] et [ASSIS] sont passés à l'arrière-plan. La définition du sens 4 de *stand* du *Collins COBUILD* ajoute cependant un sème [UPRIGHT] : *If something such as a building or a piece of furniture stands somewhere, it is in that position, and is upright* [...]. Une locomotive est bien un objet de grande hauteur par rapport à un homme et donc doté de verticalité, comme le montre l'exemple suivant :

(12) A lanky X towers above the rails at Nantes.

mais c'est également un objet oblong et, comme pour le nom *house* des deux exemples du *MEDAL*, les deux verbes sont possibles.

Il est toujours délicat de raisonner en invoquant une différence entre le conceptuel et le lexical, mais le fait que *sit*, et non pas seulement *stand* qui colloque avec les noms de véhicules, apparaisse dans le corpus pour référer au stationnement d'une locomotive nous semble être un exemple frappant d'indétermination lexicale : devant une photographie de locomotive immobile, il est impossible de prévoir lequel des deux lexèmes figurera dans la légende si l'énonciateur choisit un verbe de position.

3.2. Mouvement

Les phrases-noyaux comportent souvent un contenu référentiel dynamique, exprimant le mouvement de la locomotive ou du train. On sait que l'anglais est une langue qui, à l'opposé du français, décrit la trajectoire dans une particule ou un groupe adverbial et le type de mouvement dans le verbe lexical, l'opération nécessaire lors de la traduction étant dénommée *chassé-croisé* (Vinay & Darbelnet, 1977). Le verbe d'événement n'étant pas contraint d'exprimer la trajectoire, il est disponible pour exprimer avec une grande précision n'importe quelle facette du mouvement. Pour les seules constructions intransitives (V ou V + syntagme prépositionnel ou particule adverbiale), pas moins de 43 verbes différents sont ainsi utilisés, présentés ci-dessous par nombres d'occurrences décroissants :

10	storm	glide
7	drift	hammer
	roll	hurtle
5	pull	loom
	blast	nudge
4	struggle	plug pop
	work	pound
3	race	run
	roar	sail
	thunder	scamper
2	chug	scuttle
	churn	shoot
	ease	smoke
	hurry	snake
	rumble	steam
	stride	stride
1	barrel	strut
	charge	swing
	climb	trundle
	coast	wheel
	fly	whistle

Le plus fréquent est *storm*, avec dix occurrences, dont l'emploi correspond de toute évidence au besoin de souligner l'aspect menaçant, bruyant et spectaculaire d'une machine en plein effort. Parmi les autres verbes du champ sonore, on trouve *blast* (5 occurrences), *roar* (3), *thunder* (3), *chug* (2), *rumble* (2), *hammer*, *pound*, *pop*, le sémantisme de plusieurs de ces verbes ayant en commun avec celui de *storm* des traits de force et de violence. Comme la photographie est incapable de représenter les sons autrement qu'indirectement, et encore pas toujours, il est clair que les légendes correspondantes ont pour fonction de suppléer à ce manque, voire d'apporter du spectaculaire. Lorsque c'est la seule manière du mouvement qui est exprimée, la variété est aussi considérable : *drift* (7), *roll* (7), *race* (3), *ease* (2), *barrel, charge, coast, fly, glide, hurtle, run, scamper, scuttle, shoot, snake, strut*. Cette variété est à contraster avec la répétitivité des photos. Il n'y a là rien d'étonnant au fond, puisqu'il s'agit d'*elegant variation*, une contrainte stylistique générale, mais il s'agit en plus d'éviter la monotonie entraînée par le caractère lié du nano-genre. Cette variété dans les verbes d'événement est très caractéristique de celui-ci. Comme cette *elegant variation* opère forcément sur des légendes successives, chacune constituant un (nano-)texte, on peut considérer que l'ensemble des légendes d'un recueil est un super-texte.

Une démarche onomasiologique plus large est également possible. La liste reproduite en Annexe contient toutes les phrases-noyaux référant d'une manière ou d'une autre à l'ascension d'une rampe. Les segments concernés y sont soulignés. Cette liste permet de constater à quel point l'information visuelle des photographies est sensationnalisée, la montée étant souvent décrite comme un combat. Ceci montre bien en quoi le nano-genre légendes de photographies de trains n'a rien de technique ou industriel.

Dans le domaine d'analyse difficile de la phraséologie variable, Stubbs (2001) présente une intéressante analyse de la construction SN – V-one's way-(SP). Il s'agit d'une construction très productive (100 par million d'occurrences). Citant Francis *et al.* (1996), Stubbs signale que le verbe dénote fréquemment la force et la violence, la malhonnêteté, l'illégalité ou la stupidité. Une telle construction qui permet de détailler la manière du mouvement en l'associant explicitement à un trajet parcouru est bien à sa place dans notre corpus, et c'est sans surprise qu'on constate que sur quatre occurrences, deux dénotent la force et l'agressivité :

> (12) An old but mighty X, the most powerful locomotive on 3ft 6in track anywhere in the world, muscles her way out of Tayside on the Glencoe-Vryheid line.

(13) Looking as if it were travelling at the speed of sound, this Rio Grande X is churning its way at all of 30 mph south from Alamosa to Antonito, Colorado.
(14) Five years later, things are under better control as X works its way toward the Springs near Peyton.
(15) On a freight train clawing its way up the grade to Blue Ridge, Garver holds a microphone to catch the sounds of the X pusher he's riding.

On notera en guise de conclusion à quel point un tel nano-genre, peu attirant au départ pour les non-amateurs du domaine concerné, se révèle une source de vocabulaire éminemment exploitable à des fins pédagogiques avec des étudiants avancés. Les nano-genres permettent de mettre en relief un lexique riche dans ses limites étroites.

4. RÉFÉRENCES

Biber, D. & AL. 1999. *Longman Grammar of Spoken and Written English*. London : Pearson Education.

Collins-COBUILD English Dictionary. Sinclair, J. (ed.-in-chief). 1995. London : Harper-Collins.

Concise Oxford Dictionary. Pearsall, J. (ed.-in-chief). 2001. Oxford : O.U.P. [10th ed.].

Flowerdew, J., Dudley-Evans, T. 2002. « Genre analysis of editorial letters to international journal contributors ». *Applied Linguistics*. 23, 4. 2002. pp. 463-489.

Francis, G., Hunston, S., Manning, E. 1996. *Grammar Patterns 1: Verbs*. London : Harper-Collins.

Giannoni, D. S. 2002. « Words of gratitude : A contrastive study of ackowledgement texts in English and Italian research articles ». *Applied Linguistics*. 23, 1, 2002. pp. 1-31.

Huddleston, R., Pullum, G.K. 2002. *The Cambridge Grammar of the English Language*. Cambridge : C.U.P.

Larreya, P., Rivière, C. 1999. *Grammaire explicative de l'anglais*. Harlow : Addison Wesley Longman.

Macmillan English Dictionary for Advanced Learners. Rundell, M. (ed.-in-chief). 2002. London : Bloomsbury.

Oxford English Dictionary (2nd ed. on CD-ROM). 1994. Oxford : O.U.P.

Quirk, R. & AL. 1985. *A Comprehensive Grammar of the English Language*. London : Longman.

Stubbs, M. 2001. *Words and Phrases : Corpus Studies of Lexical Semantics*. Oxford : Blackwell.

Swales, J. M. 1990. *Genre Analysis : English in Academic Research Settings*. Cambridge : C.U.P.

Vinay, J. P., Darbelnet, J. 1977. *Stylistique comparée du français et de l'anglais*. Paris : Didier.

5. SOURCES DU CORPUS

Albi, C., Jones, W. C. 1982. *Otto Perry : Master Railroad Photographer*. Golden, Colo. : Colorado Railroad Museum.
Boocock, C. 1986. *BR Steam in Colour, 1948-1968*. London : Ian Allan.
Classic Trains (USA)
Collias, J. G. 1972. *The Search for Steam*. Forest Park. Ill.
Del Vecchio, M. 1998. *Pictorial History of America's Railroads*. Godalming : Quadrillion.
Edmonson, H., Hand, V. 1970. *World Steam in Action*. Shepperton : Ian Allan.
Gilbertson, C. B. 1977. *Steam in Australia*. Truro : Bradford Barton.
Leitch, D. 1997. *The Best of New Zealand Steam*. Auckland : Image Publications Ltd.
Steam Days (UK)

6. ANNEXE : PHRASES-NOYAUX RÉFÉRANT À UNE MONTÉE

6.1. Expressions neutres

One of the competent Xs, X, climbs the Lickey incline with a northbound passenger working in D.

Ex-GW X climbs towards Talerddig with the up "Cambrian Coast Express".

Not long before the end of steam there, X climbs past Smallbrook Junction near Ryde with a train for the Shanklin line.

With X up front and an X shoving behind the caboose, an NYS westbound climbs the 1.1 percent grade at West Lafayette.

X heads up the 1 in 25 incline to Roa, 2.7 km further on from Blackball, on D.

Another X ascends the Bank the same month.

X on no. 168 goods to Springfield climbs the bank in the evening of D.

On D, X leads a four-engine freight up Raton Pass at 30 mph near Morley, Colorado.

In D, aging X pauses for water at Mole Creek Tank after climbing the 1 in 30 grade from Glenreagh with the Dorrigo mixed.

X and X are about to climb the famous Bethungra Spiral (Junee-Cootamundra) with No. 16 Riverina Express on D.

The morning sun glints of a Rio Grande Ski Train as its X climbs the Front Range of the Rockies at Fire Clay, Colorado, on D.

On the last weekday of BR steam in north-east England, D, a well-cleaned X passes Ryhope and commences the climb to Seaton bank, en route to Hawthorn mine with a train of empties.

X is climbing to Ridge, Tenn, shouldering 61 cars of southbound extra, while older Mallets of USRA vintage supply the booster power needed to gain the Blue Ridge Heights.

Leased X climbs through the canyon just out of the Durango yards with a 34-car freight on D.

Below, in D, another of the Xs, X, climbs the hill at Monument with 43 cars at a respectable 35 mph.

6.2. Expressions dénotant la force ou la lutte

In one of Link's best-known photos, X on the Abingdon branch mixed train works up the 3-percent grade into Green Cove, Va.

Few associate the CB&Q with mountain railroading, yet here a pair of Xs, X and X, work up the three percent grade from Dumont, South Dakota with 18 cars at just 15 mph in D on the rugged Edgemont-Deadwood branch.

The power of a X is epitomized by this view of X hammering over Ais Gill with a long southbound freight, one year before the end of steam on BR.

On a freight train clawing its way up the grade to Blue Ridge, Garver holds a microphone to catch the sounds of the X pusher he's riding.

A Lewis favorite, Pennsy's Wilkes-Barre-Philadelphia *Anthracite Express* roars up the 2½ percent grade at Barmouth, Pa., 7 miles from its Broad Street Station terminus, behind a X on D.

X struggles out of the Wabash Valley with the eastbound *Carolina Special*.

Just past the tower, X attacks the hill with Second 68.

X and X with a passenger train to New Plymouth attack the 1 in 35 gradient of the Westmere Bank out of Aramaho (Wanganui) on D.

Heavyweight power on the Midland Line: here X works hard ascending the 1 in 50 grades of the Cass Bank with goods train 150 on D.

X on no. 150 goods thunders up with coal and timber from the Coast in D.

An example of a X minus streamlining, X, as it storms up the 1 in 50 gradient near National Park with a goods train in D.

X storms up the Avoca Bank on the Midland Line with a goods train for Arthur's Pass in D.

X and X work up the four percent grade at Coxo, Colorado, with a solid train of "Gramps" tank cars bound for the Alamosa refinery in D.

X fills the upper Eagle River Valley with an eruption of sulphurous coal smoke as it works upgrade at 15 mph through Mitchell, Colorado, nearly 10,000 feet above sea level on the west side of Tennessee Pass.

In a very early action view below, X chugs upgrade near Sedalia in D.

Above, three Xs struggle up the west side of Raton with a relatively short train in D, while below X surmounts Glorieta Pass with only 34 cars during a heavy March snow in the previous year.

X struggles uphill from Milton as it nears the city end of the recently completed Corinda-Roma Street quadruplication on a suburban freight train in D.

In D, X tackles the gentle rise out of Stockrington in grand style with a 750-ton non-air load en route to Hexham.

X tackles the Mt Lofty Ranges in fine style on an eastbound goods near Sleeps Hill in D.

With the knowledge that their demise is imminent, X and X struggle up Yandiah Bank towards Booleroo centre (Gladstone-Wilmington) with No. 309 goods in D.

Class leader X thunders uphill out of Gladstone with 1750 tons of ore in tow on No. 107 goods in D.

Santa Fe X gets a roll on as it bites the grades of Cajon Pass, just east of San Bernardino, California.

Sound and fury at Summit Tunnel East signalbox on the former Lancashire and Yorkshire Railway trans-Pennine route as X blasts uphill out of the Calder Valley with a Healey Mills to Manchester freight.

An Elgin to Keith via Craigellachie Class 'H' freight, which includes a number of cattle wagons, in the charge of X, is at grips with the steep grades through Glen Fiddich on D.

Casting its shadow on the grassy embankment between Gorgie East and Craiglockhart, on the Edinburgh south-side suburban line, X of Grangemouth depot struggles up the 1 in 80 gradient with an eastbound train of coal empties destined for Leith or Niddrie Yards on the east of the city in D.

Hot and panting mightly from a lusty battle with the Blue Ridge, X pauses at Old Fort, NC, to gulp a tank of water and blow her top, figuratively

speaking, as safety pops and stack exhaust respond to the fireboy's readiness for the grades just ahead.

It's a long, upgrade pull for X on the line from Mombasa to Nairobi.

A picture of self-contained fury, Japanese National railways X blasts up the grade leading inland from the Hokkaido seaport of Hakodate. A sister 2-8-2, X, shoves hard at the rear.

Fighting upgrade on the line out of Espinal, X encounters spectacular mountain scenery near Gualanday (above) before arriving at Ibague (facing page) with the afternoon "mixto".

Northbound Teine assaults the grade leading to up to Kami-Mena in the charge of X and X.

Trailing a winter plume of steam and smoke, X storms upgrade toward Arthur's Pass.

7. NOTES

(1) Larreya et Rivière (1999:136-7) indiquent à juste titre que le féminin est utilisé en général pour les bateaux et parlent de "marquer de l'affection" à l'égard de certaines machines (automobiles, camions, etc.). Par ailleurs, on se rappellera que la reprise par le masculin était possible aux XVII^{e} et XVIII^{e} siècles, ce qui est compatible avec les lexies *merchantman* et *man of war* (*Oxford English Dictionary*).
(2) Les tests de significativité courants ne sont pas applicables sur des effectifs aussi faibles.
(3) *Sibling* est un terme de démographie, biologie, psychologie, etc.
(4) Les dictionnaires pour apprenants présentent l'intérêt d'avoir des définitions plus élaborées et un classement fréquentiel des divers sens des entrées polysémiques.
(5) C'est également le cas de l'*Oxford English Dictionary* et du *Concise Oxford Dictionary*, mais pas du *Collins-COBUILD*.

LE CORPUS ET LA PRAGMATIQUE : UNE HYPOTHÈSE SUR L'EMPLOI CONTRASTIF DE L'IMPARFAIT ET DU PASSÉ COMPOSÉ

Lidia Fraczak, Stéphanie Giron
Université de Clermont-Ferrand II, LRL

1. INTRODUCTION

Dans ce travail, nous partons de l'idée que l'étude du français parlé effectuée à partir d'un corpus informatisé de langue orale peut servir à de nombreux chercheurs, dans des domaines aussi variés que la syntaxe, la pragmatique ou encore la sociolinguistique urbaine, pour n'en citer que quelques-uns. En effet, comme le signale C. Blanche-Benveniste (Bilger, 2000), « nous assistons actuellement à une reconversion complète des études linguistiques, qui ne pourront plus se passer, d'ici peu, de la consultation des ressources de langue informatisées. (…) Comme ils sont informatisés, ils [les corpus] se prêtent facilement à des structurations internes en sous-corpus. Comme ils contiennent des données attestées, dont on peut vérifier les sources, ils engagent à faire un travail d'analyse linguistique qui ne repose pas uniquement sur l'intuition mais sur la confrontation avec des données parfois étonnantes, que la simple intuition n'aurait pas pu atteindre ».

Bien que l'expérience du Français Fondamental se soit avérée très utile − même si « les données rassemblées dans les années 1950-1960 […] ne portaient pas le nom de corpus » et que « leurs objectifs étaient différents » (*ibid.*, p. 12) −, force est de constater l'intérêt que peuvent présenter, aujourd'hui encore, les liens entre corpus de langue parlée et recherche linguistique, applicables notamment au domaine du Français Langue Etrangère. En effet, l'hypothèse présentée dans ce travail, si elle est vérifiée, pourra contribuer à aider les apprenants et les enseignants de FLE.

Dans la première partie, concernant le corpus, nous présentons tout d'abord la méthode utilisée pour le recueil des données, pour l'échantillonnage et pour la transcription, puis les intérêts que présente son utilisation.

Dans la seconde partie nous proposons une paire de valeurs distinctives « généralisée » opposant les emplois de l'imparfait à ceux du passé composé,

et nous la confrontons aux usages observés dans le corpus, en nous appuyant sur des exemples qui en sont issus.

2. LE CORPUS ALLIER

2.1. Présentation du corpus

Nous avons constitué durant l'année 2001 un corpus d'environ 100 000 mots, à partir d'enquêtes effectuées à Montluçon, Vichy et Moulins, auprès de 9 locuteurs dans chacune de ces villes (les critères de sélection de ces locuteurs seront précisés dans la partie consacrée aux enquêtes).

La durée de chaque enregistrement est de 30 minutes, dont 15 sont transcrites, selon les conventions données par l'équipe DELIC[1]. Ces enregistrements sont donc transcrits et informatisés (son et texte sont disponibles sur CD-ROM), voici leurs références :

	MONTLUCON	MOULINS	VICHY
18-30 ans niveau collège	CHOCM~10	COMMO~14	HOTEV~26
18-30 ans niveau bac	MUMON~44	FANMO~16	JOLIEV~32
18-30 ans niveau bac+3	SANDM~58	FRUIT~22	MARCH~40
30-65 ans niveau collège	MADMO~34	COIFM~12	MONOV~42
30-65 ans niveau bac	PAPMO~50	GILMO~24	QUALV~56
30-65 ans niveau bac+3	MUSIK~46	FRANM~20	JAMES~30
+ 65 ans niveau collège	MAMMO~38	MALLM~36	BIJOU~6
+ 65 ans niveau bac	FRAMO~18	INSTI~28	NINAV~48
+ 65ans niveau bac+3	BRUNMO~8	PHYMO~54	PETIV~52

Tous ces enregistrements sont de type « parole privée », c'est-à-dire qu'ils revêtent la forme d'un entretien entre le locuteur interrogé et l'enquêteur (le temps de parole étant occupé à environ 95% par le locuteur interrogé).

Les 9 corpus de Montluçon n'ont pas de thème bien défini : il s'agissait souvent pour le locuteur, par exemple, de nous raconter un fait marquant de sa vie ; en revanche, les corpus de Vichy et de Moulins ont été réalisés à partir du thème « petit commerce ».

Après cette présentation très générale de notre corpus, nous allons à présent détailler la méthode que nous avons utilisée pour effectuer les

enquêtes, tout en précisant, lorsque cela sera nécessaire, les points qui nous ont paru délicats, ou ceux dont la réalisation s'avère imparfaite.

2.2. Enquêtes

Plusieurs paramètres doivent être pris en compte lors de la constitution d'un corpus : en ce qui concerne l'enquête de terrain, le choix des locuteurs enregistrés est important si l'on veut rendre compte le plus finement possible de l'état d'une langue à un moment donné de son histoire ; les critères de sélection qui précèdent le choix des personnes interrogées permettent d'introduire diverses variables au cours de l'étude (variable géographique, niveau d'étude et âge).

Les enquêtes ont été réalisées en partie suivant les indications données par l'équipe DELIC. Lors de l'élaboration du corpus *Français de Référence*, il s'agissait d'enregistrer des locuteurs selon leur âge, réparti en 3 tranches, leur niveau d'étude, réparti également en 3 niveaux, et le type de parole enregistrée (3 types de parole : privée, publique et professionnelle). Nous avons repris les deux premiers critères, l'âge et le niveau d'étude, mais n'avons pas enregistré plusieurs types de parole : nous n'avons que des « paroles privées ». Cependant, nous avons ajouté à cet échantillonnage un critère géographique, qui apparaît dans le *Français de Référence*, mais de manière moins équilibrée (27 corpus pour Paris et sa banlieue, 49 pour le nord de la France et 55 pour le sud). En effet, nous avons enregistré 9 locuteurs dans chacune des trois villes de l'Allier, ce qui fournit à ce corpus un équilibre pouvant permettre, pour des recherches futures, une étude de la variation linguistique dans ce département, qui est situé au cœur de ce que les dialectologues nomment le *croissant*, c'est-à-dire la zone où passe la frontière entre langue d'oïl et langue d'oc, et où l'on trouve, au sud-est, un peu de franco-provençal.

Lorsqu'il y a enregistrement de discours, plusieurs méthodes sont possibles, selon l'étude que l'on veut mener à bien, car « les différents types de corpus de langue parlée se définissent généralement par les objectifs qu'ils se donnent : faciliter l'enseignement, faire de la recherche linguistique, collaborer à de plus larges recherches statistiques, participer à des recherches de sociolinguistique, etc. » (C. Blanche-Benveniste, *in* Bilger, 2000).

Notre enquête est donc « libre », si ce n'est qu'un thème est défini auparavant, ce qui permet au locuteur de parler relativement facilement. Nous entendons par « libre » le fait que l'entretien n'est pas directif : nous n'avions pas élaboré de questionnaire comme l'avaient fait les enquêteurs travaillant sur le *Corpus d'Orléans* (Biggs et Blanc, 1971). Pendant une demi-heure, le locuteur parle d'un sujet défini en préalable à l'enregistrement. Nous sommes intervenues le moins possible dans le

discours du locuteur : nous le relancions parfois en posant une question ouverte, ce qui permettait de développer un point précis ou de relancer son discours lorsqu'il y avait essoufflement et que, manifestement, le locuteur était « à court » d'idées.

D'autres paramètres se révèlent importants lors de l'enregistrement : le lieu, la position des locuteurs, l'intervention ou non de l'enquêteur, etc. Nous en présenterons quelques-uns au troisième point de cette partie. Auparavant, observons dans le détail les critères retenus pour la sélection des locuteurs, ainsi que le matériel que nous avons utilisé.

2.3. Méthode retenue pour les enquêtes

Le recueil des données est une étape non négligeable de la constitution d'un corpus : d'une part, le travail est très long, qu'il s'agisse des enquêtes de terrain ou de la transcription, et d'autre part, la qualité de l'analyse dépend essentiellement de celle du corpus ; c'est pourquoi il nous semble pertinent de présenter la méthode détaillée du travail d'enquête, qu'il s'agisse des critères de sélection des locuteurs ou des problèmes dûs, dans notre corpus, à la cohérence thématique ou ceux inhérents à tout enregistrement de discours.

Cette partie montrera donc les principales étapes de la constitution d'un corpus : dans un premier temps, nous présenterons les critères précédant le choix des locuteurs, puis nous indiquerons le matériel que nous avons utilisé pour mener à bien ces enquêtes.

2.4. Choix des critères de sélection des locuteurs

Pour être pertinemment exploitable dans divers domaines de la linguistique (en syntaxe, en pragmatique, mais aussi en sociolinguistique urbaine), le choix des locuteurs doit être opéré en fonction de plusieurs critères ; dans notre cas, nous avons utilisé les critères d'enquêtes utilisés par l'équipe DELIC lors de l'élaboration du Corpus français parlé de référence, au nombre de trois : critère géographique (3 zones urbaines de l'Allier : Vichy, Montluçon et Moulins), critère socioprofessionnel (3 niveaux d'études : collège, bac et bac + 3), critère d'âge (3 tranches d'âge : 18-30 ans, 30-65 ans et + de 65 ans). Ces trois critères peuvent servir de variables dans une étude ultérieure, puisque la variation linguistique peut être analysée d'un point de vue géographique, d'un point de vue sociolinguistique ou encore diachronique (l'hypothèse étant qu'un locuteur de plus de 65 ans n'utilisera pas nécessairement les mêmes tournures syntaxiques qu'un locuteur de 18 ans, et que, même si l'on sait que la syntaxe n'évolue pas aussi rapidement que le lexique, des tournures sont délaissées au profit d'autres, etc.).

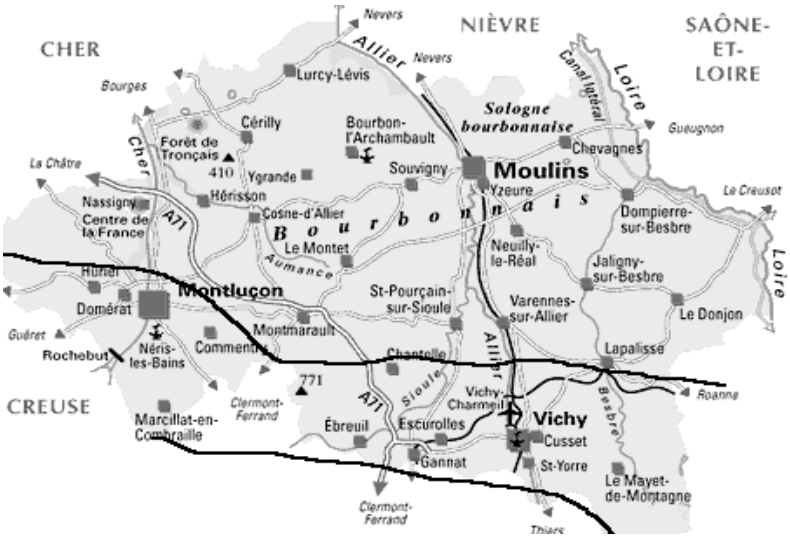
2.5. Critère géographique

Le critère géographique est celui favorisé par les dialectologues. Il permet d'observer les variations linguistiques potentielles d'un lieu à un autre si l'on choisit plusieurs lieux d'enquêtes hétérogènes (Nord/Sud, par exemple), ou, au contraire, il permet d'annuler cette variable si l'on choisit des lieux d'enquêtes situés dans une même région, dans un même département ou dans une même ville. Comme le note F. Gadet (1971) dans un article consacré aux variations sociales de la langue :

> « L'étude systématique des différences qu'elles [les variations géographiques] déterminent est du ressort de la dialectologie. Elle ne nous intéresse ici que dans la mesure où nous voyons la nécessité de choisir des sujets de même origine régionale, pour éviter l'interférence des variations linguistiques attribuables à des causes sociologiques avec celles dues à des variétés régionales. »

Cependant, il est important de noter que le lieu de notre enquête peut être considéré comme non homogène, du fait de sa situation géolinguistique : le département de l'Allier, situé en partie dans la zone que les dialectologues nomme le croissant, est à la frontière entre langue d'oïl, langue d'oc et francoprovençal. Les enquêtes étant menées en zones urbaines, nous pensons avoir limité ainsi les influences dialectales, tant il est vrai que « [un des] principaux facteurs pouvant avoir une répercussion sur la façon de parler [est] la localisation de l'habitat à l'intérieur d'une même région (ville ou campagne) » (*ibid.*).

Ainsi, le choix d'effectuer les enquêtes en zones urbaines permet d'unifier ce facteur et, par conséquent, de limiter l'importance des variations géolinguistiques que l'on aurait eues si les enquêtes avaient été menées à la campagne. Mais leur étude peut se révéler intéressante. En effet, l'emplacement géographique des trois villes à l'intérieur du département (*cf.* carte) peut permettre de comparer le français urbain standard parlé dans la zone d'oïl (Moulins) et dans la zone médiane entre oïl et oc, située dans le croissant (Montluçon, au nord, et Vichy, au sud). Ainsi, ce corpus peut permettre de vérifier l'influence dialectale sur le parler standard, ce que nous ne ferons pas dans le cadre de cette étude, car nous n'utiliserons pas ce critère. Mais, disons-le une fois encore, plus un corpus est composé de sous-types (qui sont délimités par les différentes variables retenues), plus son exploitation peut se révéler riche pour des études ultérieures.



_ **:** zone située à l'intérieur du Croissant.

Carte de l'Allier

2.6. Critère « niveau d'étude »

Ce critère comprend trois différents niveaux d'études :
- collège
- baccalauréat
- baccalauréat + 3

Il faut cependant préciser que ces trois niveaux sont indicatifs : un individu scolarisé jusqu'à la terminale et qui n'a pas passé l'épreuve du baccalauréat entre dans la catégorie « baccalauréat » ; en revanche, un individu scolarisé jusqu'en seconde entre dans la catégorie « collège » ; un individu ayant fait plus de trois années d'études supérieures entre dans la catégorie « baccalauréat + 3 ».

Selon F. Gadet (*ibid.*), le niveau d'étude est un des trois principaux indices d'appartenance d'un individu à une classe sociale, avec le salaire et la profession, indices qui, selon elle, doivent « être étudiés en bloc, sans perdre de vue que l'un de ces facteurs peut être prépondérant ». Ne retenir que le niveau d'étude pourra donc apparaître réducteur à certains sociolinguistes ; cependant, la profession de chaque locuteur apparaît dans la fiche signalétique qui lui correspond et qui figure en annexe. Ainsi, seul le salaire du locuteur n'apparaît pas.

Ce critère est intéressant pour plusieurs raisons : d'une part, il permet d'éviter de demander au locuteur (que l'on voit parfois pour la première fois) quels sont ses revenus, question qui peut bien sûr être très gênante pour

l'« enquêté », mais aussi pour l'enquêteur ; d'autre part, même si la scolarisation n'est pas le seul moyen éducatif ni le seul lieu d'apprentissage de la langue, le niveau d'étude paraît pertinent pour rendre compte de différents niveaux socioprofessionnels.

Passons maintenant à la description de ce que nous appelons « critère âge ».

2.7. Critère « âge »

Ce critère comprend trois tranches d'âge :
- 18-30 ans
- 30-65 ans
- + de 65 ans

Nous n'utilisons pas ici le terme de génération, qui est fluctuant, et qui se comprenait il y a cinquante ans comme une classe d'âge d'environ 20-25 ans, mais qui aujourd'hui n'est plus interprété de la même manière, les enfants arrivant généralement plus tard (rallongement de la durée des études, progrès de la médecine, etc.).

L'intérêt d'interroger des locuteurs de trois tranches d'âge différentes est de permettre d'avoir une vision assez précise de l'évolution d'une langue au cours du temps. En effet, un locuteur de 70 ans emploiera quelques tournures (et parfois quelques mots issus du patois) que n'emploierait pas un autre locuteur de 18 ou de 40 ans (notamment le passé surcomposé, employé par les locuteurs de plus de 65 ans, mais très rarement par les autres), comme F. Gadet le souligne (*ibid.)* :

> « Du fait qu'une langue est perpétuellement en évolution, certains traits sont parfois réalisés différemment chez les "jeunes" et chez les "vieux". Par exemple en français, il y a une disparition lente, mais progressive, de l'opposition entre /a/ et /ɑ/, entre /ɛ̃/ et /œ̃/, entre le /e/ du futur et le /ɛ/ du conditionnel, chez les jeunes locuteurs. Au niveau lexicologique, notons un emploi plus systématique de termes argotiques dans la jeunesse. »

Nous venons de passer en revue les trois critères de sélection que nous avons retenus pour nos enquêtes. Une fois cette étape achevée, l'enquêteur doit trouver les personnes y correspondant. La tâche se révèle plus ou moins aisée selon qu'il connaît ou non la localité sur laquelle il travaille. Elle peut l'être aussi à cause de tout autre phénomène, que nous n'étudierons pas ici, et qui peuvent être aussi bien liés à la timidité de l'enquêteur, à sa capacité d'« adaptation sociale », ou encore à la relation même qu'il entretient avec le locuteur qu'il interroge.

La qualité du matériel utilisé pour mener à bien ces enquêtes est aussi très importante : la qualité de l'enregistrement (qui dépend à la fois du support et du micro utilisés) garantit celle de l'écoute. Les avancées technologiques

récentes (mini-disques, format mp3) permettent de conserver un maximum de données dans un minimum d'espace, tout en garantissant une qualité stable et durable (contrairement aux bandes magnétiques). En outre, leur reproduction sur CDRom s'avère très simple pour qui possède un graveur et sait s'en servir.

2.8. Matériel utilisé

Nous avons utilisé, lors de nos enquêtes, le matériel suivant :
- un enregistreur mini-disque Sony MZ-R700 ;
- un microphone de table Sony ECM-MS907 ;
- un microphone cravate VIVANCO EM116 (utilisé lors des enquêtes dans les petits commerces, où les personnes interrogées pouvaient continuer à travailler sans être trop gênées par l'enquêteur).

Par la suite, nous avons copié les enregistrements effectués sur des mini-disques sur des CD de données au format mp3 à l'aide d'un logiciel gratuit disponible sur Internet[2] (Electronic Cosmo's MPEG suite), et ajouté à cela les transcriptions au format Word2000[3]. Les fichiers sons et textes étant référencés de manière identique, la consultation se révèle ainsi plus simple, par exemple : les corpus BIJOUV~1.doc et BIJOUV~1.mp3 sont, respectivement, la version transcrite et la version son du même enregistrement.

2.9. Enquête non directive

Contrairement à la plupart des enquêtes sociolinguistiques (notamment le Corpus d'Orléans), nous n'avons pas utilisé de questionnaire, mais avons laissé libre cours à la parole du locuteur, ce que les sociologues nomment *entretien non directif*.

Jacqueline Freyssinet-Dominjon (1997) définit ainsi l'entretien non directif :

> « L'entretien non directif, également qualifié d'entretien en profondeur, se place au plus faible degré de structuration de l'échange verbal et au plus haut degré de liberté de parole de la personne enquêtée. L'enquêteur ayant posé au départ le thème de la recherche, adopte [une] attitude […] faite d'écoute empathique et de neutralité congruente. […] L'usage de ce type d'entretien […] est recommandé […] toutes les fois que la recherche porte sur des représentations souvent latentes plutôt que sur des faits précis. »

Seul le thème était prédéfini, en l'occurrence le *petit commerce*, et présentait plusieurs avantages : d'une part, l'actualité du sujet (la décrépitude du petit commerce étant nécessairement reliée à la grande distribution, les variations sur ce thème ne manquaient pas) et d'autre part, le fait que tous les citoyens soient des consommateurs plus ou moins conscients nous a permis

d'éviter les blancs − périodes pendant lesquelles le locuteur ne sait plus quoi dire −, ou les questions trop fréquentes de l'enquêteur qui empêchent la fluidité du discours, et par là même sa réalisation la plus naturelle possible.

L'intérêt de ce type d'enquête est de favoriser la production de parole la moins « interrompue » possible, permettant ainsi de relever les formes les plus « naturelles », bien que la notion de « naturel » soit discutable lorsqu'un microphone est posé en face du locuteur.

2.10. Conditions d'enregistrement (variables contextuelles)

Pour qu'un corpus puisse être envisagé et consulté de manière homogène et univoque, les conditions d'enregistrement doivent être identiques pour chaque locuteur enregistré, afin de considérer la production orale de chaque locuteur comme égale à n'importe quelle autre. Elles font partie des *variables contextuelles* que mentionne F. Gadet (1971) :

> « [Les variables contextuelles] comprennent tout ce qui peut déterminer des différences de langage et qui n'est pas directement lié au locuteur dont on analyse le discours : la personne à laquelle il parle (et l'image qu'il se fait de cette personne), ce dont il parle, *l'emplacement dans lequel se déroule la conversation*, les rapports qui unissent les locuteurs […]. »

En effet, pour pouvoir comparer deux ou plusieurs objets, ou plus simplement pour décrire un ensemble de textes, il est nécessaire que les propriétés constitutives de chaque production orale soient en partie semblables. Nous disons *en partie* car même si, théoriquement, les conditions d'enregistrement devraient être identiques pour chaque locuteur, nous avons eu quelques difficultés d'applications pratiques, dans les cas notamment où nous enregistrions des locuteurs sur leur lieu de travail. Ceci est sans doute un des reproches qui pourront nous être adressés, nous en avons bien conscience.

Concernant notre corpus, voici les conditions d'enregistrement :

2.11. Lieu d'enregistrement

Lorsque l'enregistrement se déroule chez l'enquêté, ce dernier propose généralement la pièce habituelle de réception la plus calme (cuisine, bureau ou jardin). 24 corpus sont enregistrés ainsi.

Lorsque l'enregistrement ne se déroule pas chez l'enquêté, il est effectué soit sur son lieu de travail à l'aide d'un micro-cravate pour ne pas gêner l'accomplissement de certaines tâches (2 corpus enregistrés ainsi : James~30 et Monov~42), soit à la terrasse d'un café en zone piétonne avec le microphone « traditionnel » (Mumon~44).

2.12. Position enquêteur / enquêté

L'enquêteur fait face à l'enquêté, sauf pour James~30 et Monov~42, où la situation d'enregistrement sur le lieu de travail n'a pas permis cette position.

Ainsi, en essayant d'unifier au maximum le contexte et les conditions d'enregistrement, nous pouvons comparer différents objets ayant ce trait commun, l'idéal étant bien sûr de pouvoir établir une réelle constante à partir de ce contexte et de ces conditions d'enregistrement.

Nous avons vu jusqu'à présent la méthode d'enquête utilisée, depuis le choix des critères de sélection de l'échantillon de locuteurs jusqu'au matériel utilisé et aux conditions d'enregistrement, en passant par une présentation sommaire des lieux de l'enquête. Après cela commence un long travail d'écoute et de transcription, qui répond lui aussi à certaines règles et conventions, que nous ne présenterons pas ici, qui sont celles utilisées autrefois par le Groupe Aixois de Recherche en Syntaxe, et actuellement reprises, pour la plupart d'entre elles, par DELIC.

2.13. L'intérêt de ce corpus pour des recherches en linguistique

L'utilisation d'un corpus de données attestées en français parlé « permet une confrontation des hypothèses théoriques avec une réalité indépendante et complexe »[4]. Plus précisément, voici quelques-uns des avantages qu'une telle confrontation présente :

- les exemples sont attestés, ce qui permet au chercheur d'éviter l'écueil de l'« exemple construit », souvent jugé artificiel et qui suscite beaucoup de critiques ;
- les exemples apparaissent en contexte, ce qui permet de mieux définir les cadres énonciatifs et discursifs de leur utilisation.

Ainsi, ce corpus peut servir à des recherches en pragmatique, comme c'est le cas dans la seconde partie de ce travail, ou en syntaxe (il sert actuellement de base de travail à un inventaire et à un classement des constructions verbales), en sémantique ou encore en sociolinguistique urbaine. Mais d'autres conventions de transcriptions que celles que nous avons utilisées peuvent lui être attribuées : le linguiste phonéticien ou dialectologue, par exemple, pourra utiliser une transcription phonétique, et ainsi effectuer des recherches à partir de ce corpus, dont la localisation géographique semble particulièrement intéressante pour ces disciplines.

3. ÉTUDE DE L'EMPLOI DE L'IMPARFAIT ET DU PASSÉ COMPOSÉ

En utilisant le *corpus Allier*, nous nous intéressons à une description des différences d'emploi de deux temps verbaux du passé, l'imparfait et le passé

composé, répondant à la question : « quelle est la fonction fondamentale de chacune de ces formes » ou, autrement dit, « à quelle intention énonciative générale chacune d'elles correspond », dans les contextes où ces formes peuvent être considérées comme « concurrentes ». Il s'agit d'une description qui est, à la fois, justifiée du point de vue linguistique, et accessible aux apprenants étrangers (ainsi qu'aux enseignants eux-mêmes). Sa valeur générale doit permettre d'englober et de remplacer la variété de valeurs hétérogènes d'ordre aspectuel, discursif, stylistique ou encore extra-linguistique exprimées en termes de : « perfectif / imperfectif », « accompli / non accompli », « borné / non borné », « ponctuel / duratif », « narratif / descriptif », « premier plan / arrière-plan », effet de « ralenti », emploi « anaphorique », « habitude », etc.

Une description intéressante de la valeur générale de l'imparfait a été proposée en 1979 par Oswald Ducrot, sans qu'elle ait eu, cependant, l'impact mérité dans le domaine de la didactique du Français Langue Étrangère (FLE). Elle a été testée, avec des résultats convaincants, sur un corpus de narrations orales par Marie Labelle (1987), et reprise et développée par Jean-Claude Anscombre (1992) pour une analyse de l'opposition imparfait/passé composé. Nous nous proposons de rappeler les idées de ces linguistes et de les confronter au corpus de français parlé présenté plus haut, l'objectif étant de contribuer à une meilleure compréhension et explication des emplois de l'imparfait et du passé composé dans le domaine du FLE.

3.1. Propositions de O. Ducrot et de J.-C. Anscombre

Selon Oswald Ducrot, la différence entre l'imparfait et d'autres temps du passé réside dans « une différence de point de vue, de perspective, de centre d'intérêt » (Ducrot, 1979). Voici comment l'auteur formule, en deux propositions, sa « description générale de l'imparfait », en se servant des concepts de « thème » et de « propos »[5] :

> « Lorsqu'un énoncé est à l'imparfait, son thème est nécessairement temporel : c'est, soit une période du passé, soit, plus fréquemment, un objet ou événement considéré à l'intérieur d'une certaine période du passé (...). »
>
> « Lorsqu'un énoncé est à l'imparfait, l'état ou l'événement constituant son propos sont présentés comme des propriétés, comme des caractéristiques du thème, et qualifient celui-ci dans sa totalité. »

Ainsi, apparaît la « fonction qualificatrice » de l'imparfait : les faits décrits par les verbes à l'imparfait servent, selon Ducrot, à caractériser une période passée ou un objet ou un événement considéré pendant cette période, et « les événements présentés à l'imparfait ne constituent pas vraiment l'objet du récit » (p. 10).

J.-C. Anscombre développe, à partir de la description en deux propositions de Ducrot, une analyse contrastive des emplois de l'imparfait et du passé composé, en se servant des notions « d'espace discursif temporel » et « d'espace discursif intemporel » au lieu de « thème » et de « propos ».

Premièrement, selon cet auteur, « alors que l'imparfait nous présente des êtres temporels dont les péripéties événementielles sont conférées comme autant de propriétés constitutives, le passé composé, à l'inverse, fait intervenir des êtres intemporels, à identité permanente, dont il nous décrit l'histoire événementielle » (p. 47). Voici, en guise d'illustration, comment l'auteur commente les deux exemples suivants :

La France s'appelait autrefois la Gaule.
La France s'est appelée autrefois la Gaule.

> « La présence de l'imparfait dans le premier énoncé fait que l'espace discursif considéré est temporel. Il s'agit de "la France, autrefois", dont on dit que le nom est "la Gaule". Dans le second énoncé, à l'inverse, le passé composé introduit un espace discursif cette fois intemporel : "la France", dont on affirme qu'autrefois, elle avait pour nom "la Gaule" » (p. 48).

Deuxièmement, comme cela est d'ailleurs déjà suggéré dans le premier point, à la différence de ce qui se passe avec l'imparfait, « on n'exprime pas une propriété intrinsèque à l'aide du passé composé » (p. 51). Ainsi, un énoncé à l'imparfait, comme le premier exemple ci-dessous, est un « énoncé-propriété », tandis qu'un énoncé au passé composé, comme le deuxième exemple, est un « énoncé-événement ».

Enfant, Marie était maladroite.
Toute sa vie, Marie a été maladroite.

Selon l'auteur, le premier exemple « parle de Marie enfant (espace discursif temporel), et lui attribue la maladresse comme caractéristique » tandis que le second « s'intéresse à la maladresse de Marie (espace discursif intemporel), dont il déclare qu'elle accompagne la vie de Marie » (p. 51).

3.2. Confrontation avec des exemples du corpus.

Nous allons à présent vérifier dans quelle mesure les propositions de ces linguistes, illustrées dans leurs articles respectifs par des exemples fabriqués ainsi que, chez O. Ducrot, par quelques extraits littéraires, se défendent vis-à-vis d'énoncés attestés du français parlé, et peuvent être utiles pour le domaine de la didactique du FLE. Nous allons d'abord analyser trois fragments contenant des verbes à l'imparfait et, ensuite, trois autres avec des verbes au passé composé.

Nous nous intéresserons plus loin aux « schémas pragmatico-discursifs » auxquels les formes de l'imparfait et du passé composé participent, ce qui nous permettra de mieux ancrer les fonctions retenues de ces formes.

Imparfait

Observons les trois exemples suivants, provenant de deux enregistrements :

> 1) « quand les enfants étaient plus grands et qu'ils étaient à l'école euh bon (...) il partait le mardi (...) et il rentrait le vendredi donc c'était moi qui m'occupais de de tout ce qui pouvait se passer euh au cours de l'année scolaire » (FRAMO~18 : 10,8 – 10,12)
>
> 2) « nous étions trois ou quatre euh grandes bijouteries euh malheureusement aujourd'hui euh nous restons que quelques-uns » (BIJOUV~6 : 2,12 – 2,14)
>
> 3) « on fabriquait euh trois quatre bagues par semaine euh à la main alors que maintenant euh en l'an deux mille euh si on v- on vend des petites bagues d'un gramme cinquante deux grammes » (BIJOUV~6 : 8,3 – 8,7)

On peut se demander, dans un premier temps, si le « thème » ou ce qui est au centre d'intérêt de ces énoncés correspond, comme le postule Ducrot, à une période passée ou à un objet considéré pendant cette période. Il s'agit sans aucun doute, pour chacun de ces exemples, d'une période passée, mais on ne dirait pas pour autant que c'est cette « période » elle-même qui constitue l'aspect central de l'énoncé. Ce que l'on décrit est quelque chose de plus précis, sans que ce soit, nous semble-t-il, directement le sujet de la proposition (ou un autre « objet ») : c'est, dans nos exemples, une condition ou activité passée du sujet, dont la mention sert à caractériser une situation passée. On peut citer la remarque de M. Labelle au sujet des propositions à l'imparfait dans son corpus de narrations, qui est tout à fait en accord avec nos propres observations : « elles ne racontent pas à proprement parler, mais elles décrivent les composantes d'une situation telle qu'elle est vécue par le narrateur » (p. 18). Ainsi, nous retenons la définition suivante de la fonction de l'imparfait :

> **L'intention énonciative liée à l'emploi de l'imparfait consiste à caractériser une situation passée ; les faits exprimés par les expressions verbales sont intéressants en tant qu'ils permettent cette caractérisation.**

Dans l'exemple 1, l'énonciateur décrit sa vie familiale à une certaine époque ; les faits relatés au moyen des prédicats à l'imparfait sont au service de cette fonction descriptive. Dans les exemples 2 et 3, la situation passée, à la description de laquelle participent les verbes à l'imparfait, est explicitement opposée à la situation actuelle, décrite avec des verbes au présent.

Notre formulation plus haut reste entièrement compatible avec les propositions de O. Ducrot et J.-C. Anscombre contenues dans les seconds

points de leurs analyses respectives (concernant la « fonction qualificatrice » de l'imparfait et la distinction « énoncé-propriété »/« énoncé-événement »), tout en faisant l'économie de leurs propositions qui sont contenues dans les premiers points, portant sur la nature « nécessairement temporelle » du « thème » pour un énoncé à l'imparfait (chez Ducrot) et la distinction entre « êtres temporels » et « êtres intemporels » pour les sujets d'un énoncé respectivement à l'imparfait et au passé composé (chez Anscombre). Nous pensons, en effet, que ces propositions sont peu utiles pour notre visée applicationnelle. Nous allons revenir sur ce problème en examinant des exemples avec le passé composé dans la section suivante.

Passé composé

4) « il y a il y a eu une euh une inflation galopante et alors les gens se sont retrouvés avec une augmentation de salaire et ils ont dépensé tout le monde était heureux et puis ça a duré une quinzaine d'années » (BIJOUV~6 : 1,4 – 1,7)

5) « en mille neuf cent cinquante il y a eu une crise municipale à Montluçon euh qui a été euh tout à fait particulière » (BRUNMO~8 : 3,2 – 3,4)

6) « je me suis installé vraiment qu'en cinquante-six » (BIJOUV~6 : 7,10 – 7,11)

Nous allons donc d'abord reconsidérer la proposition de J.-C. Anscombre selon laquelle le passé composé présenterait des « êtres intemporels », c'est-à-dire des êtres « à identité permanente » (par exemple « Marie »), à la différence de l'imparfait qui, lui, présenterait des « êtres temporels » (par exemple « Marie enfant »). Il nous semble que les exemples 4 et 5 ci-dessus ne permettent pas d'appuyer cette distinction. En effet, même si on pourrait accorder aux « gens », à « tout le monde » (exemple 4) ainsi qu'à la personne qui se cache derrière « je » (exemple 6) un certain statut « intemporel », de la même manière que le fait l'auteur pour les sujets comme « la France » et « Marie », cela paraît moins facile pour les entités correspondant à « inflation galopante », « ça » et « crise municipale » dans les fragments : « il y a eu une inflation galopante », « ça a duré une quinzaine d'années » (exemple 4) et « il y a eu une crise municipale », « qui a été tout à fait particulière » (exemple 5).

En dehors de nos réserves quant à la justesse « linguistique » de cette partie de la description, ce qui nous importe surtout c'est le problème de son utilité pour les apprenants étrangers. Nous pensons qu'une explication de l'emploi d'une forme en termes de « comment elle présente quelque chose » ou, autrement dit, du point de vue de l'effet de sens auquel elle peut contribuer, n'est pas une explication opératoire dans le domaine du FLE. On

peut douter, en effet, de l'efficacité d'une règle qui indiquerait aux apprenants d'employer l'imparfait ou le passé composé pour présenter quelqu'un ou quelque chose comme un « être temporel » ou un comme « être intemporel ». Nous sommes de l'avis qu'il est préférable de formuler à cette fin une explication en termes « pragmatiques », c'est-à-dire en tenant compte des intentions discursives de l'énonciateur dans un contexte donné : « de quoi veut-il parler » (en l'occurrence d'une situation ou d'un fait) et « pour quoi faire » (pour montrer comment c'était autrefois, pour comparer une situation passée à la présente, pour expliquer ce qui s'est passé, ce qui a amené le changement, etc.).

Ainsi, nous ne retenons, encore une fois, que ces propositions de O. Ducrot et J.-C. Anscombre qui permettent de faire la distinction entre les énoncés qui caractérisent une situation et les énoncés qui caractérisent un fait. Ayant formulé une valeur pragmatique générale de l'imparfait dans la section précédente, nous le faisons ci-dessous pour le passé composé :

> **L'intention énonciative liée à l'emploi du passé composé consiste à caractériser des faits passés ; les faits exprimés par les expressions verbales sont donc intéressants en eux-mêmes.**

On peut développer cette formulation, en précisant que le passé composé est employé pour parler d'un fait (« action », « état »...) qui intéresse l'énonciateur concernant sa nature (c'est-à-dire « ce qui s'est passé ») ou/et une de ses caractéristiques particulières, comme, par exemple, la durée (cf. la dernière forme de l'exemple 4), le moment d'occurrence (cf. la première forme de l'exemple 5 et l'exemple 6) ou tout autre aspect, comme le caractère « particulier » de la crise dont il est question dans l'exemple 5.

3.3. Schémas pragmatico-discursifs.

Il est intéressant d'analyser les différents types de contextes, que nous appelons « schémas pragmatico-discursifs » dans lesquels se retrouvent les formes de l'imparfait et du passé composé. Ce type d'analyse permet, comme nous l'avons déjà mentionné plus haut, de mieux comprendre et expliquer à quelles intentions énonciatives correspond l'emploi de chacune de ces formes.

Ainsi, notre exemple 1 plus haut illustre le schéma que l'on peut qualifier de « descriptif ». Il s'agit pour l'énonciateur de décrire sa situation familiale à une époque passée. Cette situation n'est pas explicitement comparée à la situation présente, comme c'est le cas dans les exemples 2 et 3, où une telle comparaison est mise en œuvre : il s'agit de comparer les situations passée et présente dans le domaine de la bijouterie à Vichy en général (exemple 2) et pour une bijouterie particulière (exemple 3).

Dans un autre schéma représenté au sein du corpus, on exprime la transition entre les situations passée et présente au moyen d'expressions verbales au passé composé, comme dans les exemples 7 et 8 ci-dessous.

> 7) « nous étions euh l'élite des bijoutiers de Vichy enfin malheureusement maintenant ça a changé » (BIJOUV~6 : 3,5 – 3,7)

> 8) « en moins euh c'est au mois de juillet et août il y avait en permanence euh cent cinquante mille personnes à Vichy euh aujourd'hui euh à la Compagnie Fermière par exemple pour ceux qui viennent prendre leurs euh leurs bains leurs soins euh on arrive à onze mille seulement alors c'est une chute terrible qui est due à beaucoup d'événements je pense que euh ça ça ch- les gens ont changé la Sécurité Sociale a il fut un temps qui avait interdit euh enfin qui qui s'était interdit de payer les cures de les rembourser alors ça a f- aux petites gens et ils n'ont pas pu venir à leurs frais et les certains anciens euh qui étaient riches euh ont même quitté Vichy parce que les hôtels n'ont pas toujours été à leur goûts (...) et c'est tout ça qui a contribué ce que Vichy perde beaucoup de clients » (BIJOUV~6 : 9,7 – 10,8)

Dans l'exemple 7, on explique que maintenant ce n'est pas comme avant, car « ça a changé » : la transition entre les deux situations est explicitement signalée du fait d'utiliser le verbe « changer ». Dans l'exemple 8, l'énonciateur annonce explicitement que les événements qu'il va mentionner (étant, pour certains d'entre eux, liés par une relation causale) sont intéressants car ils sont responsables du changement de situation : « c'est une chute terrible qui est due à beaucoup d'événements ». Les exemples 7 et 8 peuvent être qualifiés de « narrativo-descriptifs », ainsi que l'exemple 9 ci-dessous, dans lequel la transition exprimée par les expressions au passé composé (le fait d'engager des ouvriers) concerne, non pas une situation passée et une situation présente, comme dans les deux exemples précédents, mais deux situations passées consécutives : celle qui se caractérise par trop de travail pour pas assez de personnel, et celle où quatre personnes travaillent dans l'atelier.

> 9) « en cinquante-trois cinquante-quatre on travaillait énormément on réparait beaucoup de montres euh on f- on avait euh quinze euh vingt réparations de montre à faire par jour et donc euh ce qui m'a amené à prendre un ouvrier puis deux puis trois dans les années qui ont suivi et on avait un atelier où nous étions quatre à travailler la bijouterie comprise » (BIJOUV~6 : 7,11 – 7,16)

Enfin, l'exemple 10 ci-dessous, où l'énonciateur raconte son parcours professionnel, illustre un autre schéma encore, que l'on qualifiera de « narratif ».

10) « ensuite bon ben j'y *ai* on *a pas* j'*ai pas pu* continuer les contrats puisque on (n') *avait* droit qu'à quatre contrats donc après je *suis rentrée* à faire la plonge à la à la Sagem et de tout ça ça *m'a emmenée* à travailler pendant deux ans à la Sagem euh à dans l'usine /je, j'*ai*/ *travaillé* deux ans à faire les Canal Plus » (MADMO~34 : 3,2 – 3,9)

Ce schéma convoque principalement des formes verbales au passé composé ; les éventuelles formes à l'imparfait (ici « on n'avait droit qu'à quatre contrats ») caractérisent une situation n'étant pas en relation du type « situation ancienne -> événement(s) -> situation nouvelle » avec un ou des événements décrits au passé composé, comme c'est le cas dans le schéma que nous avons appelé « narrativo-descriptifs ». La situation caractérisée par l'expression à l'imparfait dans l'exemple 10 n'a pas été modifiée par les faits décrits au passé composé.

4. CONCLUSION

Notre analyse d'exemples provenant du corpus nous a permis, à partir des propositions précédentes de O. Ducrot et de J.-C. Anscombre, de définir les valeurs pragmatiques de base de l'imparfait et du passé composé. Nous croyons que les formulations proposées sont à la fois, bien fondées du point de vue linguistique, et suffisamment simples d'accès pour pouvoir être exploitées dans le domaine de l'enseignement-apprentissage du Français Langue Étrangère. Les « schémas pragmatico-discursifs » impliquant ces formes, que nous avons pu observer au sein du corpus, jouent un rôle important dans la démonstration et dans l'illustration des valeurs préconisées. Les différents fragments du corpus mettant en œuvre ces schémas, outre l'intérêt qu'ils représentent pour un travail de recherche, pourraient servir tels quels en classe de FLE, pour des activités combinant la compréhension orale de la parole spontanée et la réflexion sur les fonctions de l'imparfait et du passé composé.

Concernant la « généralité » de la valeur pragmatique préconisée pour chacune de ces formes, elle pourra le mieux être prouvée par une démonstration la situant par rapport aux valeurs données par d'autres auteurs. Le but d'une telle analyse, que nous exposerons ailleurs, consistera à montrer que les différentes valeurs, de nature sémantique, discursive ou stylistique, peuvent être ramenées à la paire de valeurs pragmatiques que nous avons formulée ici.

4. RÉFÉRENCES

Anscombre, J.-C. 1992. « Imparfait et passé composé : des forts en thème/propos ». *L'information grammaticale*. 55. p. 43-53.

Biggs, P. et Blanc, M. 1971. « L'enquête sociolinguistique sur le français parlé à Orléans ». *Le français dans le monde*. N° 85. décembre 1971. p. 16-25.

Bilger, M. (ed.) 2000. *Corpus. Méthodologie et applications linguistiques*. Paris : Honoré Champion et Presses Universitaires de Perpignan.
Ducrot, O. 1979. « L'imparfait en français ». *Linguistische Berichte*. 60. p. 1-23.
Freyssinet-Dominjon, J. 1997. *Méthodes de recherche en sciences sociales*. Paris : Montchrestien, Coll. « AES ».
Gadet, F. 1971. « Recherches récentes sur les variations sociales de la langue ». *Langue Française*. N° 9. *Linguistique et société*. Paris : Larousse. p. 74-81.
Giron, S. 2004. *Corpus Allier*. Transcription de 7h30 de français parlé dans l'Allier (03) en zones urbaines. Thèse de doctorat de l'Université Blaise Pascal. Clermont-Ferrand.
Labelle, M. 1987. « L'utilisation des temps du passé dans les narrations françaises : le passé composé, l'imparfait et le présent historique ». *Revue Romane*. 22 (1). p. 3-29.

5. NOTES

(1) DELIC, 2002, Projet « Corpus français parlé de référence », rapport final, non-publié.
« L'équipe DELIC (Jeune Equipe JE 2328) s'intéresse à la description des structures morphosyntaxiques et lexicales en français, en synchronie comme en diachronie. Sa méthodologie repose sur l'utilisation systématique de grands corpus oraux et écrits, à l'aide d'outils informatiques appropriés (concordanciers, étiqueteurs, outils de gestion et de navigation, etc.). » (*Cf.* http://www.up.univ-mrs.fr/delic/index.html)
(2) http://p2pfr.free.fr/win/convertisseur/MPEGSuite.exe
(3) © Microsoft Corporation, juin 1999.
(4) D. Willems, *in* Bilger 2000, p. 153.
(5) Le caractère peu opératoire de ces concepts a cependant été relevé par C.-J. Anscombre (1992) et par J.-P. Desclés (2000).

OPPOSITION ENTRE *DE* ET *DES* DEVANT LES NOMS PRÉCÉDÉS D'ÉPITHÈTE EN FRANÇAIS : PORTÉE DU « POIDS »

Itsuko Fujimura, Mitsumi Uchida, Hiroshi Nakao
Université de Nagoya, Université féminine d'Osaka, Université d'Aïchi

1. INTRODUCTION

Lorsque le nom commun au pluriel est précédé d'un adjectif épithète, faut-il employer comme article *de* au lieu de *des* (*des conditions* > *de bonnes conditions*, *des chiens* > *de petits chiens*) ? Bien que cette règle soit toujours enregistrée dans la grammaire française et qu'elle soit enseignée dans les classes de français, on sait bien que « ce point n'embarrasse pas les seuls étrangers ; il fait question aussi pour bien des Français ; c'est une des difficultés de notre syntaxe » (Le Bidois & Le Bidois, 1967).

Cette étude a deux objectifs : d'abord de donner une description la plus exhaustive possible de cette alternance au point de vue aussi bien stylistique qu'historique, basée sur l'analyse statistique de corpus de très grand taille, et ensuite, d'examiner la portée de la notion du degré de « poids » que nous avons proposée dans Fujimura *et al.* (2004), en tant qu'hypothèse expliquant de nombreux facteurs qui conditionnent ce phénomène.

Les facteurs relatifs à cette opposition que nous avons déjà relevés dans le travail cité sont : le degré de collocation entre l'épithète et le nom (question de mot composé), la caractéristique lexicale des adjectifs, la fonction informative de l'épithète et le renforcement adverbial de celle-ci. Nous y avons démontré que ces facteurs sont abordables avec la notion du degré du « poids », définie comme « importance informative » : *de* est préféré quand l'épithète est « moins légère », alors que *des* l'est quand cette dernière est « plus légère ». Dans ce présent travail par contre, nous examinerons d'abord les quatre autres facteurs concernant ce phénomène : période dans l'histoire, genre de texte (niveau de langue), liaison phonique et genre grammatical (longueur phonétique de l'épithète). Les deux premiers

portent sur le côté sociolinguistique du phénomène et les deux derniers, sur le côté linguistique ou phonétique. Nous donnerons ensuite, suivant le résultat de l'examen des facteurs, une modification à la définition de la notion de « poids ». Le « poids » n'y sera plus « l'importance informative », mais « l'importance » tout court sur plusieurs échelles y compris l'échelle phonétique. On pourra dire que le « poids phonétique » est aussi pertinent dans le choix entre *de* et *des* : plus l'épithète est phonétiquement court (ou « léger »), plus *des* est choisi, plus elle est longue (ou « moins léger »), plus *de* est préféré. Cette extension de la notion de « poids » n'est d'ailleurs pas arbitraire, mais bien au contraire, c'est l'origine même de la notion de « poids », qui a été proposée en premier lieu par des phonéticiens en tant que critère explicatif de l'ordre des mots (Hagège 1982 ; Abeillé & Godard 2000 ; Arnold *et al.* 2000 ; Wasow 1997).

2. CORPUS ET DONNÉES

Cette étude est fondée sur une base de données constituée d'environ 14 500 exemples de « *de*/*des* + (adverbe) + ADJ + NOM (au pluriel) » établie à partir des corpus mentionnés dans le tableau 1. Ce sont toujours des séquences qui se trouvaient à suivre un lemme des verbes : *avoir, être* et *prendre*, et les prépositions : *à, avec, dans, par, pour* et *sur*, les adjectifs étant *anciens*, *beaux, bons, grands, gros, jolis, mauvais, nouveaux, petits* et *vieux* (ou leur forme féminine)[1]. Nous y avons aussi inclus des exemples qui étaient sujets des verbes : *avoir, être et prendre*[2].

Nous avons également consulté deux corpus du français parlé : Corpus Allier (Giron, 2001) et *Choix de textes de français parlé, 36 extraits* (Blanche-Benveniste *et al*, 2002)[3].

Nous présenterons les résultats de l'enquête statistique dans ce qui suit, et prouverons que le degré du « poids » joue un rôle déterminant dans le choix entre l'article *de* et *des*.

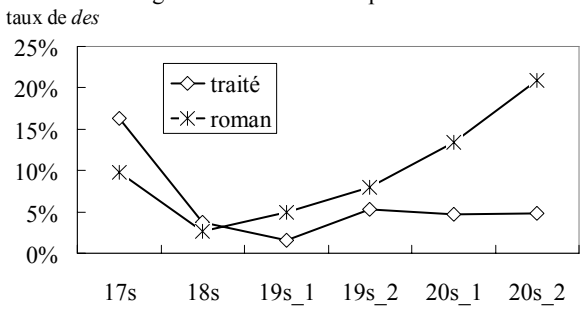
Période	Genre	Détail	Année	Nombre de mots (milliers)
20e s-2	Hansard	Les interventions en français dans le débat au Parlement Canadien, parlé officiel	1986-1988	3 300
	journal	*Le Monde* (Wordbanks*Online*)	1997 et 2001	13 000
		Libération (Wordbanks*Online*)	1992-1993	1 500
	revue	*Actuel* (Wordbanks*Online*)	1990-1992	2 000
		Marie Claire (Wordbanks*Online*)	1990-1993	3 100
	Forum de discussions (FD)	323 Forums de discussions abonnés, écrit non officiel sur les réseaux électroniques	10/2000	21 200
	roman	Frantext base catégorisée	1951-2000	15 600
	traité ou essai			11 900
20e s-1	roman	Frantext base catégorisée	1901-1950	20 600
	traité ou essai			13 500
19e s-2	roman	Frantext base catégorisée	1851-1900	14 500
	traité ou essai			4 400
19e s-1	roman	Frantext base non-catégorisée	1801-1850	12 600
	traité ou essai			8 100
18e s	roman	Frantext base non-catégorisée	1701-1800	12 000
	traité ou essai			11 200
17e s	roman	Frantext base non-catégorisée	1601-1700	3 800
	traité ou essai			7 000

Tableau 1 : Corpus

3. FACTEURS RELATIFS AU CHOIX ENTRE L'ARTICLE *DE* ET *DES*

3.1. Période

Figure 1: Taux de *des* et période



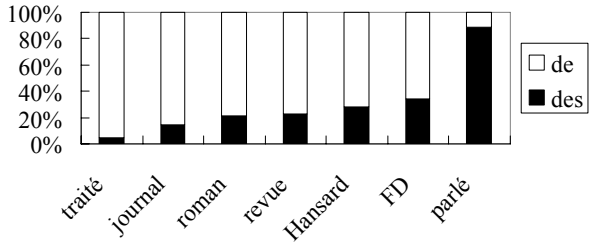
La figure 1[4] montre le changement chronologique dans le choix de l'article *des* devant les « ADJ + NOM » dans les genres du traité ou de l'essai et dans les romans de Frantext. Nous pouvons d'abord y remarquer que l'alternance entre *de* et *des* est observable depuis le tout début de l'histoire de cette règle. Par exemple, dans les traités datés du 17e siècle, le taux de *des* est 16,31 %, tandis que celui de *de*, 83,69 %. La première remarque à cet effet a été faite par Vaugelas, *Remarques sur la langue française* (1647)[5]. La lecture de ce texte, l'analyse de nos données ainsi que l'étude bibliographique (Bourciez 1976 ; Beaulieux 1927 etc.) nous amènent à supposer qu'il s'agissait principalement des dialectes régionaux à cette époque. Le choix de l'article dépendait largement des auteurs de texte au début du 17e siècle. Par exemple, Olivier de Serre, agronome né à Villeneuve-de-Berg (Ardèche) en 1539, utilise majoritairement *des* : 65 occurrences et, à l'opposé, 11 occurrences de *de* dans son ouvrage : *Le théâtre d'agriculture et mesnage des champs*[6]. En revanche, Antoine de Montchrestien, auteur du *Traicté de l'oeconomie politique*, né en 1576 à Falaise (Normandie), emploie 15 fois *de* sans aucune occurrence de *des*[7]. Après le 18e siècle on ne trouve plus de divergence aussi importante suivant les auteurs, si ce n'est la variation dépendant du niveau de langue.

3.2. Genre de texte (niveau de langue)

Nous vérifions ensuite avec la figure 2 que le niveau de langue est un des paramètres très importants de notre phénomène[8]. Cette figure indique la proportion des usages de *de* et *des* dans les textes contemporains de la 2nde moitié du 20e siècle. Parmi les corpus présentés au tableau 1, *des* est le plus utilisé dans les Forums de discussion qui sont les textes les moins officiels, et le moins dans le genre du traité ou de l'essai dont les styles sont plus soutenus. Même avec un style décontracté, l'occurrence de *de* est beaucoup

plus fréquente que celle de *des* dans tous les textes écrits. Cependant, si on porte la vue sur les textes de la langue parlée spontanée, contrairement à la discussion contrôlée comme dans le corpus Hansard, on peut remarquer que les occurrences de *des* sont beaucoup plus fréquentes que celles de *de*. Dans le Corpus Allier (Giron, 2001), le taux de *des* s'élève à 92,5 % (*de* : 5 / *des* : 62) et dans Blanche-Benveniste *et al.* (2002), il est de 86 % (*de* : 4 / *des* : 25)[9]. Nous précisons que dans ces deux derniers corpus parlés dont la taille est beaucoup plus restreinte que les autres, nous avons recueilli, contrairement aux autres, les occurrences de « *de /des* + ADJ + NOM (au pluriel) » dans tous les contextes pour avoir un nombre suffisant d'exemples, tout en recherchant les mêmes adjectifs : *ancien, beau, bon, grand, gros, joli, mauvais, petit* et *vieux* (ou leur formes féminines).

Figure 2 : *de / des* dans les textes contemporains



Il est intéressant de comparer les conditions de présence des occurrences exceptionnelles de *des* dans les traités ou essais, textes académiques, et celles de *de* dans les textes parlés. D'abord on va traiter la question des adjectifs : 23 occurrences de *des* sur 30 dans les textes académiques sont accompagnées de l'adjectif petit (les noms qui le suivent sont : bassins, bâtiments, chefs, choses, femmes, gains, garçons, gueules, jobs, nœuds, parties, pédoncules, rougeurs, spasmes, tables, toits, traces, travaux, valeurs) contre 617 de *de*. Nous avons affirmé, dans Fujimura et al. (2004), que l'affinité extrêmement forte entre *des* et petit(e)s s'explique par la notion de « poids lexical ». Petit est un adjectif lexicalement très « léger » en même temps que *des* est l'article employé avec une épithète discursivement ou lexicalement « légère ». C'est la raison pour laquelle petit(e)s apparaît très souvent avec *des*.

La deuxième observation concerne la question de la liaison : dans les traités, 29 occurrences de *des* sur 30 se produisent avec les noms qui commencent par une consonne, c'est-à-dire dans la condition où la liaison phonique ne se produit pas. La tendance de cooccurrence entre l'usage de *des* et les séquences « sans liaison » est statistiquement significative dans ce corpus ($X2 = 6,22$; $dl = 1$; $p < .05$; SANS LIAISON : *des* : 29 / *de* : 478 ; AVEC LIAISON : *des* : 1 / *de* : 139). La seule exception de *des* suivi de

« ADJ + LIAISON + NOM » est la suivante :

> Il est douteux que la seule familiarisation avec des grands auteurs suffise pour réussir cet apprentissage. (Dumazedier & Ripert, Loisir et culture 1965).

Nous considérons que la liaison est un des facteurs qui conditionnent le choix de l'article. On va revenir ci-après à l'examen de ce problème.

La troisième observation repose sur l'accentuation de l'épithète. On peut observer clairement dans les corpus parlés que l'on utilise plus souvent *de* quand l'adjectif est accentué au moyen d'adverbes notamment avec très ou bien par redoublement (X2 = 37,02 ; dl = 1 ; p < .01 ; AVEC ACCENTUATION : *des* : 2 /*de* : 9 ; SANS ACCENTUATION : *des* : 87 / *de* : 9) comme les suivants :

1. on a de gros gros problèmes cliniques par rapport à leur sida (B-B, p. 99, l.62).
2. je te fais de gros gros bisous (B-B, p. 115)
3. parce que euh /de, Ø/ toute façon depuis l'Antiquité il y a toujours eu des bijoux et de très beaux bijoux (BIJOUV~6, 6.6-6.7, Corpus Allier)
4. des fois on ne sait pas comment l'aborder parce que il a pris de très mauvaises habitudes (JAMES~30, 5,7-5,8, Corpus Allier)

Nous pouvons expliquer ce phénomène de la même manière que dans Fujimura et al. (2004) : « On peut facilement postuler que la présence d'un adverbe ajoute une information de plus ; on peut aussi estimer qu'elle rend la phrase plus complexe ou "plus lourde", ... L'épithète devenant "moins légère", la possibilité du choix de de augmente. » (p. 464).

Bien que le facteur du niveau de langue joue un rôle indubitablement très important dans notre question, il n'est tout de même qu'un des facteurs parmi les autres. Les données nous enseignent que le critère de « poids » est crucial dans tous les genres de textes, depuis les textes académiques formels jusqu'aux conversations familières en situation relâchée.

4. LIAISON PHONIQUE

Nous passons à la question phonétique. Quand le nom commence par une voyelle ou un *h* muet[10], autrement dit dans la condition où l'on fait la liaison « obligatoire » entre l'adjectif et le nom, il y a une tendance claire à éviter *des* comme l'indique le tableau 2 (X^2 = 86,30 ; dl = 1 ; p < .01). On choisit plutôt *de grands yeux*, *de nouvelles aventures* au lieu de *des grands yeux, des nouvelles aventures.* Cette condition « phonétique » qui n'a jamais été prise

en compte nous paraît soutenir notre idée primordiale que le choix de l'article *de* ou *des* a un rapport étroit avec la question de l'ordre des mots (Fujimura *et al.*, 2004). Il est connu que dans de nombreuses langues du monde, les séquences plus longues sont mises plus en arrière dans les énoncés (Hagège, 1992), et cette caractéristique de « longueur » est un constituant par excellence de l'échelle de « poids » (Arnold *et al.* 2000 ; Wasow 1997). On peut donc facilement prévoir que, quand l'épithète est plus longue avec une liaison phonique, son « poids » devenant plus lourd, l'usage de *de* tend à augmenter. Au contraire, si l'épithète est courte, donc « légère » sans liaison, l'usage de *des* est poussé.

	de	*des*
sans liaison	8719	1477
avec liaison	2879	254

Tableau 2 : Liaison et choix entre de et des

Nous montrons avec la figure 3 que les séquences, où l'adjectif et le nom sont sans liaison ont une affinité plus forte avec *des* que celles avec liaison, et cela dans toute l'histoire depuis le début de ce phénomène ; la ligne qui correspond aux « sans liaison » est constamment au-dessus de celle des groupes « avec liaison »[11]. Nous pouvons également voir la même tendance dans la figure 4 qui signale le taux de *des* dans les textes contemporains de la 2nde moitié du 20e siècle : les séquences sans liaison (ex. *grands projets* (*des :* 5 / *de :* 15), *bonnes nouvelles* (des : 3 /de : 14) ont plus de chances d'apparaître avec *des* que celles avec liaison (ex. *grandes ambitions* (*des* : 0 / *de* : 6), *bonnes intentions* (*des* : 0 / *de :* 11))[12].

Figure 3: taux de *des* et liaison dans toute l'histoire (traités et romans)

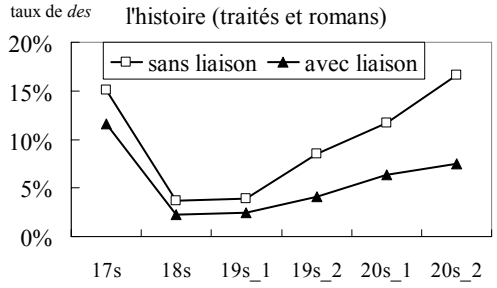
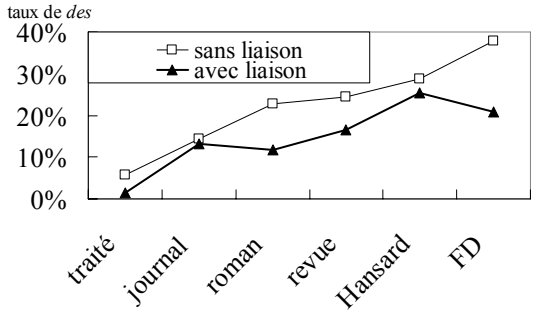


Figure 4: taux de *des* et liaison dans tous les genres (2e moitié du 20e siècle)



Ce phénomène n'est cependant pas si facile à interpréter qu'il n'y paraît.

On se demande d'abord quel est le statut du facteur phonique dans les textes écrits, puisque les données dans les figures 3 et 4 sont toutes recueillies à partir de textes rédigés, excepté Hansard qui est la transcription des discussions au Parlement Canadien. Il faut aussi se demander ensuite si la « règle » pour la liaison entraîne réellement et constamment une production phonique (dans le langage externe ou interne) chez les auteurs des textes. Nous n'avons pas de moyen, dans les textes écrits, de vérifier si des liaisons se font véritablement suivant la « règle ». Nous affirmerons tout de même avec les deux arguments qui suivent que le critère proposé est approprié, en tant que facteur du « poids phonétique », même dans les textes écrits.

Le premier argument repose sur les descriptions de phonologues ou phonéticiens. D'après Delattre (1966), « la liaison est obligatoire entre l'adjectif antéposé et le nom qui le suit dans la conversation soignée courante » (p. 43-44) et, en général, « la liaison se fait davantage au pluriel qu'au singulier. »[13] (p. 41). On peut dire que nos séquences « ADJ + NOM » au pluriel sont les cas où la règle pour la liaison est la plus strictement appliquée, par exemple dans *de grands arbres*, *de belles histoires*.

Toutefois Delattre dit également : « la liaison se fait plus fréquemment après voyelle qu'après consonne » (p. 59) et cite *mes petites amies* [meptitami] sans [z], pour l'opposer à *mes petits amis* [meptizami] avec [z] (p. 58)[14]. Cette description nous a amenés à vérifier l'enregistrement sonore du Corpus Allier. Le deuxième argument provient du résultat de cet examen. Le résultat est clair, même si le nombre d'exemples est limité dans ce corpus : la liaison en [z] est produite sans aucune exception, que ce soit après voyelle ou après consonne, par exemple dans *belles entreprises*, *bonnes années, bons élèves, petits avantages, petits avatars, petits épiciers, très mauvaises habitudes, bons élèves, longues études, jeunes amoureux*. Suivant Delattre (966), « la liaison se fait d'autant moins que le style est plus familier » (p. 40). On peux donc naturellement dire que les auteurs des textes écrits, dont le style est en principe plus soutenu que celui du parlé, effectuent la liaison en [z] dans tous les cas suivant la « règle » dans leur langage interne au cours de la rédaction de texte.

Cependant, la liaison n'est pas seulement une question phonique. Il s'agit aussi de la fonction sémantico-fonctionnelle parce que la présence de la liaison en [z] signifie le pluriel, comme on le voit dans *petits épiciers* ou *mauvaises habitudes,* alors que l'absence de la liaison n'indique rien sur le nombre comme dans *petites villes* ou *mauvais garçons*. On peut donc aussi dire que les séquences avec liaison en [z] sont moins « légères » que celles sans liaison, parce que les premières sont plus informatives que les dernières.

5. GENRE GRAMMATICAL DE L'ADJECTIF (LONGUEUR DU MOT)

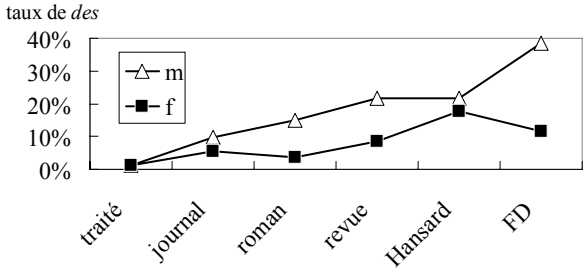
Nous examinons en dernier lieu le genre grammatical, au point de vue phonique. Afin de bien mener cette étude, nous avions besoin de résoudre au préalable le problème de l'hétérogénéité de répartition entre le genre grammatical et ses référents[15]. Le genre grammatical n'est pas sémantiquement neutre, car, par exemple, les noms humains au pluriel sont le plus souvent masculins, tandis que les deux tiers des noms abstraits sont féminins. Pour ne traiter que l'aspect formel du genre grammatical et éliminer autant que possible des facteurs sémantiques, bref, pour assurer une homogénéité sémantique dans les données, nous avons analysé, dans les textes contemporains, 1 791 exemples constitués de 153 noms abstraits trouvés dans le *Dictionnaire fondamental de la langue française*. Le jugement de catégorisation pour « abstrait » a été fait par nous-même. Ce sont des noms exprimant une notion ou un fait plus ou moins abstrait, telle que *raisons, problèmes, projets, travaux, gestes, changements, incidents, conditions, aventures, raisons, connaissances, surprises, baisses* etc., qui s'opposent aux noms dont le référent est un objet ou un humain avec un contour concret plus ou moins précis[16].

	de	*des*
m (= forme courte)	504	141
f (= forme longue)	1054	92

Tableau 3 : Genre grammatical et choix entre de et des

Le tableau 3 indique le résultat sous ces conditions : l'adjectif au masculin a tendance à apparaître avec *des* et celui au féminin, avec *de* (X^2 = 69,78 ; dl = 1 ; p < .01)[17]. On peut interpréter ce résultat de la même manière que le précédent : plus l'adjectif est long avec sa forme féminine, plus *de* est choisi ; plus l'adjectif est court avec la forme masculine, plus *des* est choisi.

Figure 5: m/f et genre de texte
(2e moitié du 20e siècle, noms abstraits)



Nous pouvons constater cette même tendance dans la figure 5 : le taux de *des* est constamment plus élevé avec la forme masculine qu'avec la féminine

dans tous les genres des textes contemporains[18]. La comparaison de la figure 5 et de la figure 4 précédente nous permet de faire remarquer que la forme des courbes graphiques est presque identique : sur les deux figures, les lignes s'approchent dans les journaux et dans Hansard, et elles s'écartent dans les Forums de discussion (ces faits ne sont tout de même pas faciles à interpréter). Ce qui signifierait que ces deux critères, liaison et genre grammatical, ont la même nature. Ils concernent la longueur phonique des mots. En d'autres termes, c'est la question du « poids phonétique ».

6. CONCLUSION

Nous avons examiné un à un quatre facteurs pour le choix de l'article *de* et *des* qui n'avaient pas été traités dans Fujimura *et al.* (2004), soit la période dans l'histoire, le genre de texte, la liaison phonique, et le genre grammatical. Le schéma suivant présente les résultats de l'examen :

Période dans l'histoire

Genre de texte

des < = > moins soutenu <---------- > plus soutenu < = > *de*

Poids de l'épithète

des < = > plus léger <---------- > moins léger < = > *de*

lexical

petit > .. *grand* ..*beau*, .. *nouveau* > .. *nombreux, excellent* > .....*(carré, présidentiel, ..)*

tout > *très* > .....*(véritablement, politiquement, ..)*

discursif

moins informatif > plus informatif
constituant d'un mot composé > épithète pleine
sans adverbe > avec adverbe

phonétique

sans liaison > avec liaison
genre masculin > genre féminin

Nous avons donné une modification à la notion du « poids », dont la définition précédente était l'importance informative des mots. La notion du « poids » a maintenant une portée plus large avec l'idée phonétique comme

partie intégrante de cette nouvelle définition. L'épithète est plus « lourde », donc donne plus de possibilité au *de*, lorsqu'elle est plus longue, plus complexe, plus informative ou plus prédicative, et elle est plus « légère », donc amène plus de *des* dans le cas contraire. Il faut enfin remarquer que ce sont des critères exactement identiques à ceux proposés pour résoudre la question de l'ordre des mots.

Le critère de « période dans l'histoire » n'est pas à intégrer dans la notion de « poids »[19]. Par contre, le « genre de texte » ne serait pas impossible à envisager en tant que facteur du « poids ». On peut dire en général que plus le style du texte est soutenu, plus l'information qui s'y trouve est dense ; plus il est relâché, moins elle l'est. L'information transmise par l'épithète antéposée pourrait donc être plus abondante ou plus prédicative dans les textes du registre soutenu que dans la conversation relâchée. Il ne serait donc pas possible de rendre compte de la tendance à l'occurrence très élevée de *de* dans les textes académiques, non seulement du point de vue purement stylistique mais aussi du point de vue de la quantité d'informations que véhicule l'épithète.

7. RÉFÉRENCES

Abeillé, A. & Godard, D. 2000. « French Word Order and Lexical Weight », *in* R. Borsley (ed.). *The Nature and Function of Syntactic Categories, Syntax and Semantics*. 32. p. 325-360.

Arnold, J., Wasow, Th., Losongco, A. et Ginstrom, R. 2000. « Heaviness vs. Newness : The effects of complexity and information structure on constituent ordering ». Language 76. p. 28-55.

Blanche-Benveniste, Cl., *et al.* (éd.) 2002. *Choix de textes de français parlé, 36 extraits*. Honoré Champion.

Beaulieux, Ch. 1927. *Histoire de l'orthographe française, tome premier, formation de l'orthographe, des origines au milieu du XVIe siècle*. Librairie Ancienne Honoré Champion.

Bourciez, E. & J. 1976. *Phonétique française, étude historique*. Klincksieck.

Delattre, P. 1966. *Studies in French and comparative phonetics, selected papers in French and English*. Mouton.

Encrevé, P. 1988. *La liaison avec et sans enchaînement, Phonologie tridimensionnelle et usages du français*. Seuil.

Fujimura, I. Uchida, M. & Nakao, H. 2004. « *De* vs *des* devant les noms précédés d'épithète en français : le problème de *petit* », *in Le Poids des mots vol 1, (Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles)*. Presses Universitaires de Louvain. p. 456-467.

Giron, S. 2001. Corpus Allier.

Gougenheim, G. 1958. *Dictionnaire fondamental de la langue française*. Nouv. éd. revue et augmentée. Didier.

Hagège, Cl. 1982. *La structure des langues*. PUF.

Le Bidois G. & Le Bidois R. 1967. *Syntaxe du français moderne*, *Tome 1*. Éd. A. Picard.

Manguin, J. 2004. « L'évolution en français de l'adjectif épithète vers la postposition : réalité syntaxique ou trompe-l'œil lexical ? » *in Le Poids des mots vol 2, (Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles)*. Presses Universitaires de Louvain. p. 753-763.

Vaugelas, Cl. F. de 1647. *Remarques sur la langue française : utiles à ceux qui veulent bien parler et bien escrire.*

Wasow, Th. 1997. « Remarks on Grammatical Weight », *Language Variation and Change 9*. p. 81-105.

8. NOTES

(1) L'adjectif *jeune* que nous avons traité dans le travail précédent n'est pas pris en compte ici, puisque son comportement combinatoire est bien différent des autres.

(2) Nous avons manuellement examiné tous les exemples pour ne conserver que ceux qui sont appropriés à notre objectif : « ART + (ADV) + ADJ + NOM » (au pluriel). Voir les détails dans Fujimura *et al,* 2004 : p. 456. Comme nous avons restreint les contextes afin de n'obtenir autant que possible que des exemples nécessaires et suffisants pour notre but, notre base de données ne représente en fait qu'une sous-classe du phénomène. Ce qui n'entraîne cependant pas de conséquence erronée pour notre but. En effet, l'examen des données nous enseigne que la fonction grammaticale du SN dans la phrase ne joue pas un rôle pertinent pour le choix entre *de* et *des*.

(3) D'après notre propre comptage, les tailles des corpus sont d'environ 89 000 et 66 000 mots respectivement.

(4) Les occurrences en nombre réel sont les suivantes :

Traité : 17s (*des* : 92 / *de* : 472), 18s (*des* : 40/ *de* : 1031), 19s_1 (*des* : 10 / *de* : 606), 19s_2 (*des* : 16 / *de* : 282), 20s_1 (*des* : 38 / *de* : 761), 20s_2 (*des* : 31 / *de :* 617)

Roman : 17s (*des* : 26 / *de* : 241), 18s (des *:* 23 / de : 853), 19s_1 (*des* : 43 / *de* : 822), 19s_2 (*des* : 103 / *de* : 1186), 20s_1 (*des* : 213 / *de* : 1381), 20s_2 (*des* : 227 / *de* : 862)

(5) Voici le texte intégral de l'article. *De*, vs. *Des* articles. Je doutois si j'en ferois une Remarque, mon dessein n'estant que d'en faire sur les choses, qui sont tous les jours en question & en dispute, mesme parmy les gens de la Cour, & nos meilleurs Escrivains. Il ne me sembloit pas que celle-cy deust estre mise en ce rang, comme en effet, il n'y a gueres de personnes qui ayent tant soit peu de soin d'apprendre à bien parler & à bien escrire, qui ne sçachent ce que je vais remarquer. Néanmoins ayant considéré, que dans la plus part des Prouinces, on y manque, & que parmy ce nombre infini d'Escrivains qui sont en France, il y en a une bonne partie, qui ne prennent pas garde, j'ay jugé cette Remarque nécessaire. Au nominatif, & à l'accusatif *de* se met devant l'adjectif, & *des* devant le substantif, par exemple on dit, *il y a d'excellens hommes*, & *il y a des hommes excellens*, *ce pays porte d'excellens hommes*, vs. *porte des hommes excellents*, & non pas *il y a des excellens hommes*, ny *il y a d'hommes excellens*, & ainsi de l'autre. C'est une reigle essentielle dans la langue, J'ay dit que c'estoit au nominatif & à l'accusatif, qu'elle avoit lieu, parce qu'au génitif & à l'ablatif, il n'en va pas ainsi. Car on dit *la gloire des excellens hommes*, & *on l'a despouillé des belles charges qu'il possedoit*. (Vaugelas, 1647, p. 331).

(6) L'usage de *des* et *de* nous semble conditionné chez cet auteur par le sémantisme des noms. *De* est utilisé lorsque la pluralité du référent n'est pas informative, par exemple avec des noms : *roues, fruits, carreaux, ciseaux, cormes, fromages, diversités, miels*.

(7) On peut supposer aussi une différence du niveau de langue.

(8) Les occurrences en nombre réel sont les suivantes : traité (*des* : 31 / *de* : 617) ; journal (*des* : 145 / *de* : 886) ; roman (*des :* 227 / *de* : 862) ; revue (*des* : 116 / *de* : 388) ; revue (*des :* 116 / *de* : 388) ; Hansard (*des :* 78 / *de* : 201) ; FD (*des* : 530 / *de* : 1009) et parlé (*des* : 87/ *de* : 9).

(9) Les exemples avec adverbe ou redoublement entre ADJ et NOM ne sont pas inclus aux chiffres. Voici tous les exemples de « *de* ADJ NOM (au pluriel) » dans les corpus parlés :
- tout le monde n'a pas forcément besoin de faire de grosses courses pour aller acheter euh un litre d'huile un pot de moutarde euh trois œufs ben (COIFM~12 (1,3-1,5), Giron (2001))
- nous sur une région comme Moulins on /n', a/ pas des grandes surfaces qui ont de gros stocks hein (COIFM~12 (15,1-15,2), Giron (2001))
- ça soit quand même par rapport à notre région qu'on ait peut-être plus de chance d'avoir de bons produits que que dans certaines grandes autres villes euh comme Lyon (FRUIT~22 (6,12-6,14), Giron (2001))
- tu vois c'est un petit peu euh un peu comme ça on a bu de bons coups là-bas (MUSIK~46 (16,9-16,10))
-les FNAC n'existaient pas les choses comme ça bon les l- les libraires étaient de petits libraires indépendants (PHYMO~54 (6,10 - 6,12), Giron (2001))
- allô maman c'est Julien ça va – passé de bonnes fêtes – bon moi ça va (B-B, p.114, l.4)
- en face de moi un tapis roulant euh qui menait directement euh à des jolies demoiselles qui mettaient de b-belles étiquettes de toutes les couleurs sur nos sur nos petits poulets (B-B, p.158, l.27)
- c'était un hôtel euh très exotique − − on nous a conduits euh à notre bungalow − l'hôtel avait euh de petits bungalows – (B-B, p.161, l.20)
- mais c'est très difficile à réconforter cette cette personne qui /est, Ø / d'ailleurs souffre physiquement énormément − et bon ben les nerfs em- l'emportent euh dans de dans de grandes discussions euh à savoir euh qu'est-ce qu'elle a fait ou qu'est-ce qu'elle a pas fait de bien − mais elle elle a du mal à nous l'expliquer on a eu énormément de mal à élucider ce problème (B-B, p.184, l.599)
- voilà donc trois types un petit peu de de de de personnes rencontrées les personnes indépendantes les les personnes dépendantes qui sont en maison et les personnes qui sont en situation précaire et qui eux qui elles bon ben ont quand même de de grosses difficultés à s'insérer dans la société actuellement (B-B, p.174, l.205)
(10) Les semi-voyelles sont souvent aspirées, comme *petits week-ends, grands yachts.*
(11) Les occurrences en nombre réel :
17s : sans liaison (*des* : 93 / *de* : 521), avec liaison (*des* : 25 / *de* : 191), 18s : sans liaison (*des :* 50/ *de* : 1302), avec liaison (*des :* 13/ *de* : 562), 19s_1 : sans liaison : (*des* : 44 / *de* : 1044), avec liaison : (*des* : 8 / *de* : 349) , 19s_2 : sans liaison (*des* : 103 / *de* : 1090), avec liaison (*des* : 15 / *de* : 338), 20s_1: sans liaison (*des* : 210 / *de* : 1603), avec liaison (*des* : 34 / *de* : 492), 20s_2 : sans liaison (*des* : 961 / *de* : 2989), avec liaison (*des* : 157 / *de* : 920)
(12) traité : sans liaison (*des* : 29 / *de* : 477), avec liaison (*des* : 2 / *de* : 139) ; journal : sans liaison (*des* : 107 / *de* : 630), avec liaison (*des* : 38 / *de* : 250) ; roman : sans liaison (*des* : 204/ *de* : 679), avec liaison (*des* : 23 / *de* : 168) ; revue : sans liaison (*des* : 100 / *de* : 293), avec liaison (*des* : 15 / *de* : 74) ; Hansard : sans liaison (*des* : 61/ *de* : 151), avec liaison (*des* : 17 / *de* : 50) ; FD : sans liaison (*des* : 460 / *de* : 759), avec liaison (*des* : 62 / *de* : 239)
(13) Delattre suggère d'autre part que la liaison est conditionnée par la longueur des mots (1962, p. 41) : « *La liaison se fait d'autant plus que le premier des deux mots est plus court :* En un jour ; depuis un jour, Très utile ; extrêmement utile. » On constate ici que le « poids » phonétique semble aussi pertinent dans l'application de la liaison.
(14) Encrevé (1988, p. 234) cite les même exemples.
(15) De même, nous n'avons pas pris en compte l'adjectif *joli(e)s* qui est phonétiquement épicène.
(16) Le facteur sémantique est aussi pertinent pour le choix de l'article. *Des* est employé avec des noms indiquant les êtres humains plutôt qu'avec des noms abstraits. *Des* est aussi utilisé plus souvent pour les noms indiquant des objets concrets et artificiels, que pour les parties du corps qui sont plurales de nature, comme *les yeux*, *les dents*, etc. Nous ne pouvons pas en discuter ici plus en détail à cause de l'espace limité de l'article.

(17) Les occurrences dans la totalité des données sont les suivants : m (*des* : 1072, *de :* 5987) ; f (*des* ; 641, *de* ; 5414). La tendance de cooccurrence entre le masculin et *des* et entre le féminin et *de* est significative ($X^2 = 60{,}73$; dl = 1 ; $p < .01$).
(18) traité : m (*des* : 1 / *de* : 71), f (*des* : 2 / *de* : 134) ; journal : m (*des* : 12 / *de* : 85), f (*des* : 11 / *de* : 208) ; roman : m (*des* : 16/ *de* : 87), f (*des* : 3 / *de* : 85) ; revue: m (*des* : 8/ *de* : 23), f (*des* : 4 / *de* : 56) ; Hansard : m (*des* : 8/ *de* : 24), f (*des* : 12 / *de* : 59) ; FD : m (*des* : 89/ *de* : 141), f (*des* : 37 / *de* : 238)
(19) Mais il serait intéressant d'étudier de plus près des faits comme dans Manguin (2004).

STRATÉGIE DE CONSULTATION DE CORPUS ORAUX TRANSCRITS : PISTES MÉTHODOLOGIQUES POUR L'EXPLORATION D'UN CORPUS THÉMATIQUE À HAUT RENDEMENT[1]

Nathalie Gasiglia
Université de Lille III − SILEX, CNRS UMR 8528[2]

1. CONTEXTE DE TRAVAIL

Je reprends ici la notion de « corpus à haut rendement »[3] afin de mettre en perspective l'évaluation des différentes procédures d'exploration de ces ressources documentaires[4]. Des corpus de ce type sont plus intéressants à constituer, pour certaines visées lexicographiques, que des corpus plus aléatoires[5], pour lesquels le coût de consultation serait inutilement élevé, dans la mesure où les documents intégrés et explorés ne répondraient pas à des besoins très précisément exprimés, analysés et critérisés : si le corpus consulté est de taille importante, l'avalanche de données extraites pourra noyer le lexicographe qui les consulte, sans garantir ni l'exhaustivité ni la qualité (les documents trouvés sur le Web, par exemple, sont, selon les sites consultés, d'une tenue linguistique discutable) ; s'il est de taille plus réduite, le consultant sera moins submergé par la quantité d'extractions produites, mais la part de silence risque d'augmenter au-delà de ce qui peut être acceptable, c'est-à-dire au-delà du seuil de rentabilité qui fait qu'un lexicographe accepte de passer du temps à consulter un corpus, alors que le travail de rédaction presse. J'ai argumenté ailleurs[6] pour que ces corpus à haut rendement se concrétisent sous la forme de corpus plurithématiques d'une technicité « bien tempérée »[7], conçus comme des agglomérats progressifs de corpus monothématiques consistants. Dans le cadre de l'expérience relatée ici, le corpus est constitué de transcriptions de dialogues radiodiffusés pour la documentation lexicale d'un thème spécifique, le football, et, pour mettre en avant ses qualités, je me placerai dans la situation de lexicographes mettant en chantier un lexique où seraient explicités les comportements linguistiques particuliers, dans une langue spécialisée de grande diffusion, de mots couramment employés par ailleurs, dans la

pratique la plus courante de la langue commune. Appliquée au football, cette approche pourrait présider à l'élaboration d'un dictionnaire de langue spécialisée massivement diffusée ou, pour un dictionnaire de langue générale, à la rédaction d'articles visant la description d'un large spectre d'emplois des items présents à la nomenclature. L'adjectif *dangereux,* par exemple, est un mot courant de la langue française dans la description duquel il convient que soient pris en compte trois facteurs : (i) l'expression d'un danger inhérent ou provoqué (avec respectivement un protagoniste mentionné – celui qui est en danger –, ou deux – celui-ci et celui qui cause le danger), (ii) le caractère physique ou non du danger, et (iii) le fait que ce danger reçoive une appréciation positive ou négative. Ils sont illustrés ci-dessous au moyen de trois expressions, dont la première relève de la langue commune, alors que les deux suivantes sont d'un usage plus restreint et entrent dans la langue spécialisée étudiée :

	Protagoniste mis en danger	**Danger physique ou non**	**Appréciation positive ou négative du danger**
*une **chute** dangereuse*	**celui qui chute** (le seul mentionné)	**physique**	**négative**
*un **tacle** dangereux*	**le joueur victime du tacle** (fait par un joueur de l'équipe adverse (*cf.* Fig. 3))	**physique**	**négative**, puisqu'il s'agit d'une irrégularité commise par un joueur faisant ainsi courir un risque à l'intégrité physique d'un joueur de l'équipe adverse
*une **action** dangereuse*	**l'équipe adverse** (de celle qui conduit l'action)	**non physique** : tactique	**positive** pour qui conduit l'action, puisqu'il s'agit d'une phase de jeu dans laquelle une équipe se trouve en situation de marquer un but contre l'équipe adverse

Même si les lexicographes sont en principe des locuteurs particulièrement attentifs, ils ne peuvent pas mobiliser instantanément, lors de la rédaction de chacun de leurs articles, tous les emplois régulièrement observés, que cette régularité soit propre à une langue spécialisée de grande diffusion, comme le football, ou observable au sein de plusieurs pratiques (par exemple les sports collectifs, de ballon ou non).

Pour évaluer selon quelles modalités il convient d'explorer un corpus à haut rendement, comme celui établi pour l'étude de l'expression en français d'expériences du football, afin (i) qu'il facilite le repérage des mots et expressions effectivement employés, (ii) qu'il donne accès à l'observation des emplois et permette, partant, d'en évaluer la régularité et d'en fournir des descriptions linguistiques les plus fines et les plus justes possibles, je vais évoquer différentes méthodes, "manuelles" ou plus "automatiques", de dépouillement, pour me concentrer en dernier lieu sur des procédures

informatiquement outillées et quelques pistes de travail explorées à ce jour. Mais cela ne sera fait qu'après avoir décrit le *Corpus foot de multiplex transcrits* dans son état actuel, et montré en quoi sa structure a été conçue pour maximiser la rentabilité de sa consultation.

2. PRÉSENTATION DU *CORPUS FOOT DE MULTIPLEX TRANSCRITS*

2.1. Contenu du corpus

Le corpus considéré a été envisagé comme un corpus à haut rendement. Il n'intègre que des transcriptions d'oral : les propos énoncés par des animateurs en studio, des reporters sur le terrain, des entraîneurs, des présidents de clubs et des joueurs durant neuf journées du championnat de France 2002-2003, retransmises sur plusieurs chaînes de radio sous forme de multiplex. Ces neuf documents audio, une fois transcrits, fournissent un corpus XMLisé (saisi avec l'éditeur de documents XMetaL et structuré au moyen d'un balisage écrit en langage XML) de 200 000 mots environ, au sein duquel sont balisés les tours de parole et les interventions simultanées (importantes parce qu'elles peuvent réorienter les propos au cours de leur développement), assortis à chaque fois de l'identité des locuteurs et de leur statut, afin que les descriptions des usages lexicaux observés puissent tenir compte des conditions d'énonciation et des types de locuteurs[8]. Enchâssés dans ces premiers éléments qui structurent les transcriptions, sont en particulier balisés (i) les prononciations déviantes observées, avec repérage de celles qui sont considérées comme pertinentes pour un travail lexicographique[9], (ii) les mots partiellement prononcés, auxquels (pour limiter le silence lors des extractions automatiques) est associée leur forme pleine quand cela s'est avéré possible de façon assurée, (iii) les interjections non codifiées[10], qui, bien que non lexicalisées, jouent un rôle dans les énoncés et, à ce titre, ne doivent pas disparaître, etc. Les conventions de transcription que nous utilisons ont été établies indépendamment de celles du DELIC (2004) et ne sont pas directement compatibles : si les principes sont le plus souvent concordants, la mise en œuvre du codage est très différente, probablement en large partie du fait que nous partions « de rien » lorsque nous avons débuté les transcriptions en langage XML, alors que le DELIC a hérité du savoir-faire et des corpus du GARS et qu'il devait poursuivre l'œuvre de celui-ci avec un souci de cohérence. Sur quelques points, nos positions divergent, vraisemblablement du fait des exploitations visées et de leurs incidences sur les objets à repérer : nous avons élaboré ces transcriptions à partir de sources particulières, des commentaires radiophoniques de matchs de football, pour mener à bien des études lexicales et, même si nous avons fait en sorte de réserver la possibilité de procéder à d'autres explorations de ce corpus, il n'est pas préparé pour

elles. Nous n'avons pas marqué les pauses qui, dans le cadre des multiplex, sont mal discernables tellement le débit de parole des commentateurs est rapide. Nous n'avons pas non plus noté les bruits de fond, aussi régulièrement présents sur les stades que peu importants pour notre objet. Nous n'avons rien anonymisé dans la mesure où il s'agit de parole publique. Et nous avons cherché à ne pas mélanger le texte de la transcription et les repérages ou commentaires de nature diverse, qui sont tous localisés dans le balisage.

2.2. Structuration et modalités de transcription du corpus

La structure du document XML produit en 2002-2003 est présentée ci-dessous avant un extrait de transcription où il sera aisé d'observer qu'elle respecte quelques principes simples : pas de ponctuation, en particulier pour ne pas générer d'interprétations sélectives voire fautives, pas de normalisation, et des transcriptions orthographiées selon l'usage courant, respectueuses des normes typographiques habituelles et de l'orthographe enregistrée dans au moins l'un des trois dictionnaires de référence (*cf.* note 10), sauf au sein des éléments destinés à notifier des variations jugées remarquables (*cf. supra* et Fig. 1, dans la DTD (Définition de Type de Document)), mais ces informations sont isolées par leur balisage, puisqu'il s'agit ici déjà d'analyse. Pour ce qui concerne la structuration en XML, l'en-tête (non développé ci-dessous) est conforme aux préconisations du Corpus Encoding Standard (CES – Ide & Véronis, 1996), mais, faute de trouver dans les propositions relevant de la TEI (Text Encoding Initiative), le jeu de balises permettant de repérer, au sein des transcriptions, les éléments dont nous jugions avoir besoin, j'ai mis en œuvre un balisage spécifique, dont les noms d'éléments et d'attributs sont proposés en anglais pour les harmoniser avec ceux de l'en-tête, et permettre plus aisément leur réemploi par d'autres éventuels utilisateurs. Dans la DTD, présentée ci-dessous sous forme indentée, les noms d'éléments sont suivis des noms d'attributs dont ces éléments sont porteurs (noms encadrés de « ^{A} »), puis, lorsque cela s'avère utile, des opérateurs de répétition (+) ou d'optionnalité (?) portant sur chaque élément considéré dans sa globalité (balises et contenu). Des commentaires assurent enfin une bonne compréhension de chaque particularité de codage mise en place dans la DTD.

```
CORPUS
    HEADER ATYPEA <!-- en-tête de corpus conforme au CES -->
    DOCUMENT
        HEADER ATYPEA <!-- en-tête de document conforme au CES -->
        TEXT
            (
            SPEAKER AIDENTITYA APOSTA +
            <!-- pour l'élément SPEAKER on spécifie, en attribut, l'identité (IDENTITY) et
            la fonction (POST) de la personne qui parle durant le tour de parole -->
                (
                données textuelles
```

OU
ENTITY [A]TYPE-OF-ENTITY[A]
<!-- les noms de joueurs, de clubs ou de stades de football, mais plus largement de personnes, de villes, d'entreprises ou d'établissements, sont repérés comme étant des entités nommées ; ce sera également le cas des titres d'émissions diverses qui pourront être mentionnés, etc. -->
(
données textuelles
OU
SPELLED-INITIALS
<!-- sigles épelés -->
OU
READ-INITIALS
<!-- acronymes, c'est-à-dire sigles dont les lettres sont lues continûment, comme dans un mot ordinaire -->

OU

REMARKABLE-PRONUNCIATION
<!-- sigle ou nom prononcé de manière remarquable et transcrit en API (*cf.* [tefese], note 9) -->
) +
OU
REMARKABLE-UTTERANCE
<!-- quand la prononciation d'un segment est remarquable, mise en correspondance de la version orthographiée et de la transcription, en Alphabet Phonétique International, de ce qui est effectivement prononcé ; ce qui est jugé remarquable peut être (i) une distorsion locale liée à une faute, par exemple d'accord (*aucun n'auront*), qui devra être décrite dans l'élément ANALYSIS-OF-REMARKABLE-UTTERANCE (verbe au pluriel alors que son sujet, *aucun*, est au singulier), (ii) une liaison fautivement présente [katrzadø] ou absente [vuave], (iii) l'emploi de mots étrangers non enregistrés par les trois dictionnaires de référence (*cf.* note 10) comme appartenant au lexique français et pour lesquel la langue d'origine devra être indiquée via l'élément ANALYSIS-OF-REMARKABLE-UTTERANCE, etc. -->
TRANSCRIPTION-OF-REMARKABLE-UTTERANCE
<!-- segment prononcé transcrit en API -->
données textuelles
SPELLING-OF-REMARKABLE-UTTERANCE
<!-- segment orthographié -->
données textuelles
ANALYSIS-OF-REMARKABLE-UTTERANCE ?
<!-- analyse ou commentaire métalinguistiques relatifs à l'énoncé jugé remarquable -->
données textuelles
OU
ACCIDENTAL-PRONUNCIATION
<!-- quand la prononciation d'un segment est accidentellement fausse, mise en correspondance de la transcription en API de ce qui est effectivement prononcé et de la version orthographiée -->
TRANSCRIPTION-OF-ACCIDENTAL-PRONUNCIATION

<!-- segment prononcé transcrit en API -->
données textuelles
SPELLING-OF-ACCIDENTAL-PRONUNCIATION
<!-- segment orthographié -->
données textuelles
OU
AMBIGUOUS-PRONUNCIATION
<!-- quand la prononciation d'un segment est ambiguë, qu'elle peut donner lieu à plusieurs transcriptions concurrentes ou que ce qui est prononcé est perçu de manière différente par le transcripteur et les relecteurs : mise en correspondance de la transcription en API de ce qui est perçu et de la ou des différentes versions orthographiées possibles avec, éventuellement, répétition de l'ensemble, si, lors des relectures des transcriptions, les relecteurs n'entendent pas la même chose et donc ne proposeraient pas la ou les mêmes transcriptions -->
(
TRANSCRIPTION-OF-AMBIGUOUS-PRONUNCIATION
<!-- segment prononcé transcrit en API -->
données textuelles
SPELLING-OF-AMBIGUOUS-PRONUNCIATION +
<!-- segment orthographié pouvant correspondre à ce qui a été entendu, avec autant de répétitions de cet élément qu'il y a de manières d'orthographier le segment prononcé -->
données textuelles
) +
OU
INDISTINCT-PRONUNCIATION
<!-- prononciation que l'on peut transcrire en API quand elle est discernable, mais que l'on n'a pas su mettre en correspondance avec un mot ni donc orthographier -->
TRANSCRIPTION-OF-INDISTINCT-PRONUNCIATION ?
<!-- transcription en API de ce qui a été effectivement prononcé quand cela est discernable -->
données textuelles
OU
SHORTENED-IN-THE-TEXT
<!-- traitement d'un mot dont la prononciation est débutée mais non achevée -->
TRANSCRIPTION-OF-REAL-PRONUNCIATION-OF-SHORTENED-WORD
<!-- transcription en API de ce qui a été effectivement prononcé -->
données textuelles
POSSIBLE-SHORTENED-WORD ?
<!-- mot débuté mais non achevé, s'il est possible de l'identifié de manière fiable -->
données textuelles
OU
UNCODIFIED-INTERJECTION
<!-- interjection pour laquelle aucune orthographe n'est proposée dans les dictionnaires, cf. note 10 -->
TRANSCRIPTION-OF-REAL-PRONUNCIATION-OF-UNCODIFIED-INTERJECTION

```
                    <!-- transcription en API de l'interjection qui a été effectivement prononcée -->
                        | données textuelles
                OU
                REMARK A TYPE-OF-REMARK A
                <!-- les remarques correspondent à des ajouts du transcripteur ou d'un relecteur, elles peuvent être de trois types : (i) des notes sur l'avancement ou les difficultés propres au travail de transcription (orthographe incertaine en particulier), (ii) des analyses métalinguistiques ou (iii) des informations référentielles rendant mieux compréhensible l'énoncé en cours -->
                    | données textuelles
                )
            OU
            MANY-SPEAKERS-SIMULTANEOUSLY +
            <!-- pour la zone de recouvrement des paroles de plusieurs locuteurs, quand les propos tenus par un locuteur sont audibles mais superposés à ceux tenus par au moins un autre locuteur qui débute sa prise de parole -->
                | SPEAKER A IDENTITY A A POST A +
            OU
            CUT
            <!-- pour marquer, dans la transcription, les coupures liées la segmentation de l'échantillon -->
```

Figure 1 : DTD du Corpus foot de multiplex transcrits

```
<MANY-SPEAKERS-SIMULTANEOUSLY>
    <SPEAKER IDENTITY="Bernard Abbadie" POST="reporter terrain">
        <ENTITY TYPE-OF-ENTITY="joueur">
            Darcheville
        </ENTITY>
    </SPEAKER>
    <SPEAKER IDENTITY="Eric Chamoy" POST="reporter terrain">
        réduction
    </SPEAKER>
</MANY-SPEAKERS-SIMULTANEOUSLY>
<SPEAKER IDENTITY="Eric Chamoy" POST="reporter terrain">
    du score ici au stade
    <ENTITY TYPE-OF-ENTITY="stade">
        Jean Laville
    </ENTITY>
    <ACCIDENTAL-PRONUNCIATION>
        <TRANSCRIPTION-OF-ACCIDENTAL-PRONUNCIATION>
            por
        </TRANSCRIPTION-OF-ACCIDENTAL-PRONUNCIATION>
        <SPELLING-OF-ACCIDENTAL-PRONUNCIATION>
            pour
        </SPELLING-OF-ACCIDENTAL-PRONUNCIATION>
    </ACCIDENTAL-PRONUNCIATION>
    pour
    <ENTITY TYPE-OF-ENTITY="équipe">
        Créteil
    </ENTITY>
    c'est
```

```
    <ENTITY TYPE-OF-ENTITY="joueur">
        Libbra
    </ENTITY>
    parti en contre-attaque
    <ENTITY TYPE-OF-ENTITY="joueur">
        Libbra
    </ENTITY>
    qui vient de tromper euh
    <ENTITY TYPE-OF-ENTITY="joueur">
        Trivino
    </ENTITY>
    on jouait dans le temps additionnel donc quarante-septième minute deux à un pour
    <ENTITY TYPE-OF-ENTITY="équipe">
        Gueugnon
    </ENTITY>
    face à
    <ENTITY TYPE-OF-ENTITY="équipe">
        Créteil
    </ENTITY>
</SPEAKER>
<SPEAKER IDENTITY="Pierre-Louis Basse" POST="animateur studio principal">
    un petit mot à
    <ENTITY TYPE-OF-ENTITY="ville">
        Bordeaux
    </ENTITY>
</SPEAKER>
<SPEAKER IDENTITY="Bernard Abbadie" POST="reporter terrain">
    oui ballon bordelais avec
    <REMARKABLE-UTTERANCE>
        <TRANSCRIPTION-OF-REMARKABLE-UTTERANCE>
            afolo
        </TRANSCRIPTION-OF-REMARKABLE-UTTERANCE>
        <SPELLING-OF-REMARKABLE-UTTERANCE>
            affolo
        </SPELLING-OF-REMARKABLE-UTTERANCE>
        <ANALYSIS-OF-REMARKABLE-UTTERANCE>
            apocope supposée de “affolement”
        </ANALYSIS-OF-REMARKABLE-UTTERANCE>
    </REMARKABLE-UTTERANCE>
    de la défense troyenne qui se dégage malgré tout un zéro on est toujours dans le temps
    supplémentaire
</SPEAKER>
```

Figure 2 : Extrait du Corpus foot de multiplex transcrits

2.3. Améliorations visées

Nous verrons, au fil des sections suivantes, qu’en l’état le *Corpus foot de multiplex transcrits* permet d’extraire des données de qualité. Mais ne pouvons-nous pas le rendre plus rentable encore à exploiter pour l’élaboration de descriptions lexicales ? Son rendement serait-il augmenté en synchronisant les transcriptions aux sources sonores ? Cette hypothèse semble hautement

plausible dans la mesure où l'option de n'introduire aucune ponctuation dans la transcription des documents primaires, principe de codage adopté à titre conservatoire pour ne pas risquer de surinterpréter les données, peut générer des ambiguïtés que le retour à la source lèverait. Nous[11] testons actuellement[12] la plus-value que présente effectivement le fait de pouvoir revenir aux documents d'origine. À cette fin, une seconde phase de transcription est actuellement en cours, elle concerne le même type de documents sources, mais est faite de manière à pouvoir consulter simultanément la source sonore et la transcription. Le logiciel Transcriber est utilisé à cette fin : il s'agit d'un éditeur XML dédié à la transcription d'émissions radiophoniques (http://www.etca.fr/CTA/gip/Projets/Transcriber/). Il est paramétré pour permettre le codage d'un certain nombre d'événements, dont seulement certains sont nécessaires dans le cadre de la constitution du corpus qui nous occupe. Il génère un document XMLisé où le premier niveau de structuration des transcriptions est de nouveau le tour de parole (prise de parole d'un locuteur ou chevauchement des propos tenus par plusieurs personnes), avec identification et caractérisation du ou des locuteur(s) impliqué(s), et où les éléments balisés dans le premier corpus sont également repérés, mais des contraintes liées à l'interface de saisie m'ont amenée à réviser leur balisage et à les intégrer dans ce que Transcriber propose comme « commentaire »[13].

Il est trop tôt pour évaluer plus concrètement les bénéfices effectifs de cet alignement[14] dans la suite de ce témoignage. Je ne vais donc plus parler que du *Corpus foot de multiplex transcrits* existant en 2002-2003, le seul sur lequel les explorations relatées ci-dessous ont effectivement porté.

3. PRÉSENTATION DES PROCÉDURES D'EXTRACTION EXPÉRIMENTÉES ET DE LEURS QUALITÉS REMARQUÉES

Trois types de procédures d'extraction d'informations ont été testés sur un échantillon d'une quinzaine de lexicographes en formation[15] : (i) au vol, lors d'une écoute attentive et renouvelée autant que souhaitée, en notant ce qui est perçu comme méritant de l'être, (ii) en lecture de transcriptions avec surlignage et prise de notes, (iii) avec un outillage informatique en élaborant des stratégies de recherche ciblées.

3.1. Relevés au fil de l'écoute

Lors du relevé au vol durant l'écoute, même en répétant et en fractionnant l'écoute autant que désiré, la quantité d'informations pertinentes entendues est telle que tout ce qui mériterait d'être noté ne l'est pas : le corpus proposé semble trop riche pour être ainsi exploité « à l'oreille » et aucun des auditeurs ne constitue un relevé identique à celui des autres à partir d'un même extrait

analysé. À cela s'ajoute une propension bien partagée à reformuler, voire à réinterpréter : les preneurs de notes lemmatisent souvent ce qui est entendu et sélectionnent les éléments du contexte qui leur semblent pertinents, produisant ainsi des résultats de qualité variable en fonction du degré de familiarité avec ces productions langagières – celles typiques des multiplex, aussi denses que difficiles d'écoute, du fait des variations de débit et de qualité d'articulation des locuteurs, des intonations très marquées, etc. Ces différences de sensibilité d'écoute ne sont pas surprenantes, mais elles amoindrissent le profit fait de l'exploitation d'un corpus à haut rendement. Toutefois, il est intéressant de noter que, au sein de la synthèse de leurs relevés, un consensus s'est dégagé entre tous les auditeurs scripteurs pour sélectionner les mêmes objets, voire pour envisager de combler les « trous » de cette liste compilée en prenant conscience, à la lecture d'un élément repéré, du fait qu'ils avaient entendu mais non relevé une expression.

Il semble donc que si ce type de collecte est réalisé individuellement et par des personnes non déjà exercées à cette pratique, ni habituées à ce type d'expression, son produit peut se révéler relativement aléatoire, ce qui ne serait handicapant que si, dans le cadre d'un projet éditorial, un tel relevé était mené selon cette méthode et par un seul lexicographe, mais il présente l'avantage de baigner les scripteurs dans le « parler » à décrire.

3.2. Extractions au fil de la lecture

En relevé au fil de la lecture, alors que les propos sont transcrits, donc mieux accessibles puisque imprimés, ce qui permet plus aisément de revenir sur un passage (pour confirmer ou infirmer une impression mémorielle), et que les preneurs de notes les plus néophytes en football ne sont plus déroutés par le flot des paroles entendues, la différence de qualité des relevés produits reste marquée avec, en particulier, le pointage d'expressions qui n'ont pas toujours la représentativité escomptée pour ce type de lexique – ce sont éventuellement des emplois survenus dans un contexte spécifique qui a donné lieu à une construction certes remarquable, mais conjoncturellement motivée –, alors que d'autres, bien que répondant à un patron caractéristique, sont tellement régulièrement rencontrées, dans les transcriptions lues, qu'elles ne sont pas repérées comme étant significativement plus fréquentes ou différentes de celles observables dans l'ensemble de la langue commune. Par exemple, si je m'en tiens à la description du jeu et que j'exclus des constructions du type remonter au score, la construction remonter X se réalise avec deux objets remarquables : la construction remonter le ballon, qui décrit une action offensive, est courante dans le corpus exploré mais peu vraisemblable en langue commune (entendue hors langues spécialisées de grande diffusion relatives aux jeux de ballons) ; par contre, la construction remonter le terrain, qui décrit un déplacement dans le cadre d'une action offensive, n'apparaît que

rarement dans le corpus mais présente un patron régulier de construction en langue commune[16] et, à ce titre, ne devrait pas être prioritairement relevée. Contrairement à nos attentes, au sein des relevés effectués, nous trouvons régulièrement la construction remonter le terrain, alors que remonter le ballon est marginalement présent.

Il semble donc que, dans ce type d'exercice, le lecteur scripteur, bien qu'il soit attentif, est trop immergé dans l'expression d'expériences footballistiques pour en percevoir nettement les particularités les moins spectaculaires.

3.3. Explorations outillées

Pour ce qui est du dépouillement du *Corpus foot de multiplex transcrits*, dans sa version transcrite sur support électronique, au moyen d'outils informatiques (par exemple des concordanciers) accessibles à tout lexicographe[17] travaillant sur des textes pour lesquels nous disposons d'informations morphosyntaxiques associées aux mots[18], les résultats obtenus sont éminemment liés aux stratégies de recherche mises en place. Si ces outils permettent éventuellement de viser une certaine exhaustivité tant du point de vue des occurrences à relever que des phénomènes à observer, ils réclament une réflexion préalable sur les moyens d'optimiser les investigations. Manipulés trop mécaniquement, ils font courir le risque, plus massivement peut-être encore que des relevés manuels au fil de l'écoute ou de la lecture, de générer une part de silence importante[19] ou de ne pas pallier pleinement le manque de culture spécialisée des lexicographes, qui n'ont pas de représentation des situations que ce vocabulaire technique permet de nommer. Le recours aux corpus devrait leur permettre de mettre en relation les faits décrits et les termes employés pour les décrire, mais ce n'est pas aussi net quand les segments textuels extraits sont trop courts pour être effectivement informatifs ou que, le balisage XML n'indiquant pas à quelle équipe appartiennent les joueurs mentionnés[20], des erreurs d'interprétation sont rendues possibles par défaut d'information : dans une séquence comme Pagis à l'entrée de la surface de réparation qui peut peut-être décaler Santos, seul le fait que les deux joueurs Pagis et Santos appartiennent à la même équipe permet de comprendre que le verbe décaler[21] signifie « faire une passe à un coéquipier démarqué » plutôt que, par exemple, « passer au-delà d'un adversaire en conservant le ballon ».

À partir de ces premières observations, je vais consacrer la section suivante à la présentation du protocole expérimenté pour mener des investigations aussi automatisées que possible et générer des extractions dont la consultation présente un bon rendement informatif pour les lexicographes, même profanes en football.

4. QUELLES STRATÉGIES D'EXTRACTIONS AUTOMATIQUES D'INFORMATIONS POUR QUELS RÉSULTATS ?

4.1. Des extractions simples

Certaines extractions sont courantes, ce sont les listes de lemmes et de collocations[22] qui peuvent éventuellement être produites avec mention de leurs fréquences relatives, mais pour lesquelles cette dernière information sera à interpréter avec prudence, (i) parce que l'expression en français d'expériences du football ne se limite pas (Corbin à paraître ; Gasiglia 2004) au vocabulaire technique que ces listes mettent particulièrement en valeur et que, même pour les unités lexicales les plus terminologiques, la fréquence, pour un lexique spécialisé qu'il est raisonnable de considérer comme fini et épuisable dans les limites d'un volume de format honnête, n'est pas un critère de sélection, (ii) parce que la fréquence mentionnée est seulement représentative de ce que contient le corpus exploré, ce qui nous renvoie aux questions classiques de représentativité des corpus, et surtout (iii) parce que ces listes sont générées par des programmes parfois déroutants, qui peuvent ne pas identifier les unités polylexicales pour ce qu'elles sont et les découper en « mots simples », qui repèrent parfois de potentiels collocateurs sans tenir compte de la ponctuation[23], dont la présence invalidera leur réunion au sein de collocations, voire sans traiter les apostrophes donc en considérant qu'une chaîne de caractères comme l'arbitre constitue une unité lexicale et qu'elle est distincte de l'item arbitre[24]. Des solutions palliatives sont exploitables par des outils d'exploration développés pour traiter les données de manière linguistiquement plus adéquate, outils qui ne travaillent pas seulement au niveau des chaînes de caractères et/ou avec des évaluations statistiques, mais intègrent des ressources lexicales bien fournies, voire, dans certains cas, des systèmes de tri des collocations repérées distinguant celles construites sur un patron régulier, pour lesquelles une analyse syntaxico-sémantique peut être élaborée automatiquement, de celles qui ne sont pas compositionnelles. Je ne développerai pas plus ici ces évaluations de logiciels, elles sont marginales pour ce témoignage.

Au-delà de l'élaboration de la nomenclature, ce qui est attendu du corpus est une mise à disposition d'exemples plus typiques que ceux qu'un lexicographe aurait construits même si, étant sortis des contextes qui leur donnent tout leur sens, ils demandent à être partiellement reformulés pour être rendus autosuffisants et plus fluides (Corbin, à paraître). C'est par exemple le cas pour cet énoncé extrait du *Corpus foot de multiplex transcrits* :

Utaka parti euh lancé par euh Bakari qui venait de rentrer euh très bon ballon dans l'espace il a couru une bonne trentaine de mètres avec le ballon il a fixé le jeune Jody Viviani qui est rentré tout à l'heure après la blessure de Rudy Riou et il inscrit le deuxième but lensois

qui peut, sans perte d'informations utiles à la compréhension du contexte, pour illustrer le verbe *fixer*, être reformulé ainsi :

Utaka, parti lancé par Bakari, a couru une bonne trentaine de mètres avec le ballon, il a fixé le jeune Jody Viviani et il inscrit le deuxième but lensois.

Le balisage mis en place dans le *Corpus foot de multiplex transcrits* permet également d'accéder à des données spécifiques comme (i) des onomatopées ou interjections non encore répertoriées qui, particulières aux énoncés de ce type ou non, peuvent enrichir les nomenclatures des ouvrages qui leur sont dédiés, (ii) des prononciations supposées déviantes, déviances qui pourront être validées ou infirmées en quantifiant leur régularité (*cf*. la prédominance possible de [tefese] sur [teɛfse], *cf*. note 9), etc.

Enfin, d'autres extractions peuvent être conduites pour avancer dans la connaissance de la structuration conceptuelle de la langue spécialisée de diffusion massive, ici l'expression en français d'expériences du football. Elles vont m'occuper ci-après.

4.2. Des extractions stratégiquement couplées à une ontologie

Le dernier point de ce témoignage concerne des extractions qui ont été menées en suivant un protocole spécifique : pour tirer un meilleur profit des données, nous avons élaboré une ontologie générale des actions de jeu[25], afin qu'elle permette de pointer les caractéristiques remarquables des circonstances de réalisation des actions nommées, et qu'elle serve de cadre pour le stockage des premières extractions faites – en les rattachant au(x) descripteur(s) de circonstances de jeu dont elles relèvent. Cette ontologie caractérise les actions réalisées par les joueurs durant les deux mi-temps d'un match en les décrivant selon les valeurs de quatre paramètres d'analyse spécifiques à cette pratique sportive[26] : orientation du jeu (offensif ou défensif), nature du jeu (individuel ou collectif), jeu avec ou sans ballon, séquence de jeu (lors de la (re)mise en jeu, dans le déroulement du jeu ou induisant une interruption de jeu).

À partir de ces combinaisons de valeurs des quatre paramètres, la construction de l'ontologie s'est poursuivie par la mention des types d'actions de jeu sous-catégorisés, en fonction de propriétés qui leur sont propres, sous lesquels viennent prendre place les exemples lexicaux, lemmatisés après extraction en corpus. Dans l'arborescence de la Fig. 3 sont reproduites cinq combinaisons de valeurs des quatre paramètres, pertinentes pour les énoncés traités ci-après. Elles dominent chacune un type d'action de jeu, dont les propriétés particulières ne sont pas reportées pour ne pas

surcharger cet extrait d'ontologie, et des exemples lemmatisés avec renvois aux énoncés sources.

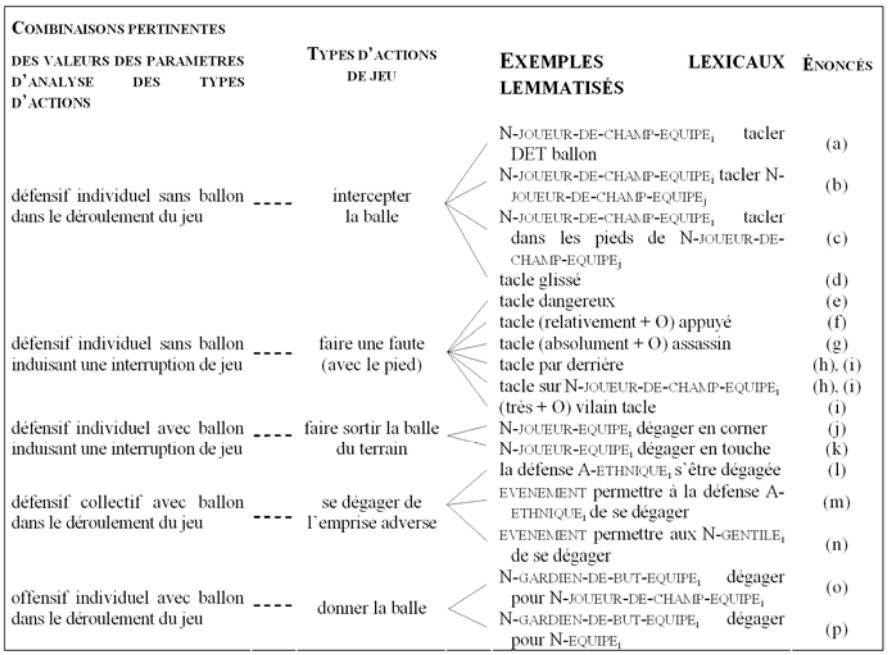

COMBINAISONS PERTINENTES
DES VALEURS DES PARAMETRES D'ANALYSE DES TYPES D'ACTIONS
TYPES D'ACTIONS DE JEU
EXEMPLES LEXICAUX LEMMATISÉS
ÉNONCÉS
défensif individuel sans ballon dans le déroulement du jeu
intercepter la balle
N-JOUEUR-DE-CHAMP-EQUIPE$_i$ tacler DET ballon (a)
N-JOUEUR-DE-CHAMP-EQUIPE$_i$ tacler N-JOUEUR-DE-CHAMP-EQUIPE$_j$ (b)
N-JOUEUR-DE-CHAMP-EQUIPE$_i$ tacler dans les pieds de N-JOUEUR-DE-CHAMP-EQUIPE$_j$ (c)
tacle glissé (d)
défensif individuel sans ballon induisant une interruption de jeu
faire une faute (avec le pied)
tacle dangereux (e)
tacle (relativement + O) appuyé (f)
tacle (absolument + O) assassin (g)
tacle par derrière (h), (i)
tacle sur N-JOUEUR-DE-CHAMP-EQUIPE$_i$ (h), (i)
(très + O) vilain tacle (i)
défensif individuel avec ballon induisant une interruption de jeu
faire sortir la balle du terrain
N-JOUEUR-EQUIPE$_i$ dégager en corner (j)
N-JOUEUR-EQUIPE$_i$ dégager en touche (k)
défensif collectif avec ballon dans le déroulement du jeu
se dégager de l'emprise adverse
la défense A-ETHNIQUE$_i$ s'être dégagée (l)
EVENEMENT permettre à la défense A-ETHNIQUE$_i$ de se dégager (m)
EVENEMENT permettre aux N-GENTILE$_i$ de se dégager (n)
offensif individuel avec ballon dans le déroulement du jeu
donner la balle
N-GARDIEN-DE-BUT-EQUIPE$_i$ dégager pour N-JOUEUR-DE-CHAMP-EQUIPE$_i$ (o)
N-GARDIEN-DE-BUT-EQUIPE$_i$ dégager pour N-EQUIPE$_i$ (p)

Figure 3 : Vue d'un fragment de l'ontologie des actions de jeu

Les premières extractions en corpus sont faites autour de mots-clés[27] (*ballon*, *joueur*, etc.) et de catégories grammaticales (les verbes, les noms, etc.) pour extraire les énoncés présentant des occurrences de certains patrons syntaxiques et y repérer des indices en fonction de divers tris effectués sur les listes de contextes produites. À ce stade, l'exploration du corpus « se nourrit elle-même », l'analyse d'une concordance conduisant à formuler un nouveau motif de recherche, puis à permettre l'établissement de classes, par exemple :
1) de verbes ou de noms synonymes pour une partie au moins de leurs emplois (à établir en fonction des patrons syntaxiques qu'ils régissent ou dont ils sont argument) et désignant des actions, d'où découle un affinement de l'ontologie initiale par l'insertion des classes de verbes (comme ceux exprimant la transmission de balle entre deux joueurs de la même équipe (Gasiglia, 2004)) et de noms comme nœuds préterminaux héritant des descripteurs de circonstances de jeu et dominant, au terme des explorations à venir, les exemples lexicaux lemmatisés, après qu'ils aient été sélectionnés au sein des concordances où leurs attestations en contexte se donnent à consulter ;
2) d'adjectifs antéposés et postposés à des noms comme par exemple *tacle*[28], en montrant qu'il y a une corrélation entre la place des adjectifs, et le fait qu'ils expriment la beauté technique du geste (*un superbe tacle*) ou qu'ils

marquent sa périllosité, son irrégularité (*un tacle dangereux*), ce qui se traduit par la mise en place de classes de qualificatifs à valeur positive ou négative, permettant de discriminer ensuite automatiquement, lors de futures explorations, les *tacles* qui sont de beaux gestes techniques (énoncés (a) à (d)) et ceux qui sont des actions à pénaliser ((e) à (i)) ;

(a) *il est repris par Zikos qui réussit à tacler ce ballon*
(b) *Meniri est revenu tacler au dernier moment Vairelles pour mettre le ballon en corner*
(c) *heureusement pour les Sochaliens Saveljic est revenu pour tacler dans les pieds de l'attaquant havrais*
(d) *c'est David Sommeil qui a sauvé dans un tacle glissé du plus bel effet*
(e) *le premier carton pour adressé à Bonnal pour un tacle dangereux*
(f) *il a été victime d'un tacle relativement appuyé de Stéphane Noro*
(g) *David Régis qui vient de faire un tacle absolument assassin*
(h) *grosse faute tacle par derrière sur Xavier Becas*
(i) *l'expulsion d'Anthony Bancarel pour un très vilain tacle par derrière sur Buengo*

3) de noms ou de syntagmes nominaux permettant de distinguer les joueurs attaquants ou défenseurs, joueurs de champ ou gardiens de but, etc., ce qui, dans le cadre de l'analyse de verbes très bien représentés comme *(se) dégager*, permet de subdiviser la liste des occurrences extraites en fonction des patrons de constructions[29] : *N-JOUEUR-EQUIPE$_i$ dégager (le ballon + la balle + Ø) (en touche + en corner + Ø)* ; *N-DEFENSE-EQUIPE$_i$ se dégager* ; *N-GARDIEN-DE-BUT-EQUIPE$_i$ dégager (le ballon + la balle + Ø) pour (N-JOUEUR-DE-CHAMP-EQUIPE$_i$ + N-EQUIPE$_i$)*.

(j) *Pitau qui lobe Penneteau sorti à sa rencontre et c'est Ferreira sur la ligne qui dégage en catastrophe en corner*
(k) *Dugarry tente de donner un ballon en profondeur et Ferreira pour Bastia parvient à dégager en touche*
(l) *deux corners consécutifs et chaque fois eh bien la défense corse s'est dégagée avec beaucoup beaucoup de fébrilité*
(m) *un attaquant troyen mais qui ne trouve personne ce qui va permettre à la défense alsacienne de se dégager*
(n) *une nouvelle fois Pagis signalé hors-jeu cela va permettre aux Havrais de se dégager*
(o) *Ferreira qui dégage pour très loin devant lui pour Piocelle*
(p) *Alexander Vencel qui peut dégager loin devant pour les Havrais*

À partir de ces regroupements syntaxiques et de leur insertion dans l'ontologie, la partition sémantique de ce verbe devient plus aisément perceptible : les formes pronominales, énoncés (l) à (n), expriment des actions

qui relèvent du jeu défensif comme les formes non pronominales, *N-JOUEUR-EQUIPE$_i$ dégager (le ballon + la balle + Ø) (en touche + en corner)*, énoncés (j) et (k), alors que celles dont l'agent est un *N-GARDIEN-DE-BUT-EQUIPE$_i$*, qui ne mentionnent pas l'objet dégagé mais ont un complément prépositionnel humain, réfèrent à des actions offensives – pour les unes, comme (o), ce complément spécifie le destinataire du dégagement (*pour N-JOUEUR-DE-CHAMP-EQUIPE$_i$*), pour les autres, comme (p), il indique quel en est le bénéficiaire (*pour N-EQUIPE$_i$*).

Les premiers dépouillements, effectués en suivant ce protocole, fournissent des données dont la qualité est tout à fait appréciable. Il semble que les analyses produites, même par les plus néophytes en football, soient mieux canalisées et, par conséquent, les descriptions du sens des mots plus conformes aux réalités du jeu qu'ils permettent de décrire. Mais il convient de rester circonspect et, même si ces premiers travaux sont encourageants, la promotion de la démarche suivie ne pourrait être envisagée qu'après de nouvelles mises à l'épreuve.

5. CONCLUSION

Ce témoignage parvient à son terme. Il a été pour moi l'occasion de revenir sur la notion de corpus à haut rendement et, de manière détaillée, de présenter les options adoptées pour la constitution de ce corpus d'oral transcrit, destiné à l'étude d'unités lexicales employées dans une langue spécialisée de grande diffusion : l'expression en français d'expériences du football. Le *Corpus foot de multiplex transcrits* est le premier grand chantier de ce type mis en place à l'université Lille III. D'autres corpus d'oral transcrit sont en cours de développement dans le cadre du projet OuRAL (*cf*. note 11). Ils sont structurés selon les mêmes principes directeurs.

Ce témoignage a également offert un espace pour rappeler, même succinctement, qu'à une époque où les outils informatiques d'exploration de corpus électroniques occupent le devant de la scène, d'autres méthodes d'exploration plus traditionnelles existent toujours, qu'il s'agisse de relevés au fil d'une écoute attentive et renouvelée ou de dépouillements au fil de la lecture, et qu'elles ont leurs qualités, même si elles s'avèrent moins adaptées à l'exploration d'un corpus à haute densité d'information comme le *Corpus foot de multiplex transcrits*. Pour lui, j'ai proposé d'employer des concordanciers en les couplant à des étiqueteurs morphosyntaxiques, d'enrichir de manière récursive les motifs de recherches à partir de ce qui est donné à voir dans des concordances initiales, et d'articuler l'analyse des données extraites avec l'ontologie des actions de jeu élaborée, à cette fin, avec Pierre et François Corbin.

Une promotion d'étudiants de la formation lilloise en lexicographie, terminographie et traitement automatique des corpus a débuté la constitution

du *Corpus foot de multiplex transcrits* et son dépouillement, focalisé sur les descriptions des actions de jeu. La qualité des premières données produites motive la poursuite de cette expérience avec d'autres, mais, au-delà, j'espère qu'elle pourra utilement nourrir la réflexion de ceux qui aspirent à rédiger ou rédigent des répertoires métalinguistiques en utilisant des corpus comme ressource documentaire.

6. RÉFÉRENCES

Bromberger, Ch. (dir.) 2002. Passions ordinaires. Football, jardinage, généalogie, concours de dictée… Paris : Hachette Littératures.

Corbin, P. 2005 (à paraître). « Des occurrences discursives aux contextualisations dictionnairiques. Éléments d'une recherche en cours sur l'expression en français d'expériences du football », in M. Heinz (éd.), *Entre définition et citation : l'exemple lexicographique dans les dictionnaires français contemporains*, Max Niemeyer Verlag, Tübingen.

Dictionnaire Hachette. Paris : Hachette Livres. 2003.

Enckel, P. & Rezeau, P. 2003. *Dictionnaire des onomatopées*. Paris : Presses Universitaires de France.

Équipe DELIC 2004. « Présentation du Corpus de référence du français parlé », in P. Cappeau (éd.), *Autour du corpus de référence du français parlé, Recherches sur le français parlé 18*, Université de Provence. p. 11-42.

Gasiglia, N. 2004. « Faire coopérer deux concordanciers-analyseurs pour optimiser les extractions en corpus », in B. Habert (dir.), *Linguistique et informatique : nouveaux défis*, Revue Française de Linguistique Appliquée, volume IX – 1, p. 45-62.

Gasiglia, N. 2005 (à paraître). « Stratégie de constitution de corpus oraux transcrits : arguments pour un corpus plurithématique à haut rendement », in G. Williams (éd.), *La linguistique de corpus en France ou en français*. Presses Universitaires de Rennes.

Habert, B., Fabre, C. & Issac, F. 1998. *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*. Paris : InterEdition.

Ide, N. & Veronis J. 1996. « Une application de la TEI aux industries de la langue : le *Corpus Encoding Standard* », Cahiers GUTenberg n° 24 (spécial TEI), juin 1996. p. 166-169.

Le Petit Robert, CD-ROM, version 2.0. Paris : Dictionnaires Le Robert / VUEF. 2001.

Le Petit Larousse, CD-ROM. Paris : Dictionnaires Larousse. 2003.

Silberztein, M., Poibeau, Th. & Balvet, A. 2001. « Intex et ses applications informatiques ». Tutoriel, *Actes de la huitième conférence TALN*, 2-5 juillet 2001, volume II. p. 145-174.

7. NOTES

(1) Merci à Pierre Corbin pour la richesse qu'il apporte à nos travaux coordonnés en formation et en recherche : cette collaboration nourrit le présent témoignage. Merci aux étudiants du DESS LTTAC (cf. note 2) promotion 2002-2003 et à François Corbin, qui ont contribué à ce que ce travail parvienne au niveau d'avancement dont je fais état. Merci à Geoffrey Williams qui a offert en 2003, pour la troisième fois, un lieu d'échanges à ceux qui

travaillent en linguistique de corpus. Merci à Pierre Corbin, de nouveau, pour ses relectures critiques de ce texte.

(2) Le DESS « Lexicographie, Terminographie et Traitement Automatique des Corpus » proposé à l'université Lille III, transformé, à partir de 2004-2005, en parcours LTTAC : parcours professionnalisant de 2e année du master « Art, Lettres, Langues et Communication », mention « Sciences du Langage », spécialité « TAL et IDL ».

(3) Je reprends ici la dénomination proposée par Pierre Corbin (à paraître : § 0, 4 et 9), option qui peut dérouter dans une introduction où on s'attendrait peut-être à ce que je lui préfère une appellation plus intuitivement compréhensible, mais qui seule semble convenir pour nommer l'objet de nos investigations sans que la qualité de la dénomination employée ne devienne un point central de la réflexion exposée. Je renvoie les lecteurs intéressés à l'analyse de Pierre Corbin développant les enjeux et implications de ce choix dénominatif.

J'utiliserai dans ce texte deux formes dénominatives, « langue spécialisée de grande diffusion » ou « massivement diffusée » et « langue commune », dont la validité pour mon propos mériterait également d'être étudiée, ce que, faute de place, je ne ferai pas ici.

(4) Dans le numéro de *RFLA* dirigé par Benoît Habert (Gasiglia 2004a, p. 48-52) et dans les actes, à paraître, des secondes *Journées de Linguistique de Corpus* de Lorient.

(5) À défaut, peut-être, de disposer d'un corpus de référence pour le français, les corpus explorés sont souvent des compilations de textes disponibles en version électronique et réunis selon des principes de commodité d'accès (les articles de tel journal, édités sur CD-ROM ou téléchargeables, voire le Web dans son ensemble, pénétrable via les moteurs offrant des recherches « plein texte » comme Google), ce qui conduit parfois à qualifier ces « corpus » d'opportunistes (Habert *& al.* 1998, p. 35).

(6) *Cf*. Gasiglia 2004, p. 50-51 et Gasiglia à paraître : section « Trois options pour la constitution de corpus pour la lexicographie ».

(7) Il s'agit de se donner les moyens d'observer les échanges langagiers particuliers qu'une large part des locuteurs d'une langue sont susceptibles de partager dans certaines situations de la vie courante présupposant une compétence technique repérée chez leurs acteurs (Bromberger (dir.), 2002) : la cuisine, le bricolage, le jardinage, un sport, une activité artistique ou manuelle, etc. Nous approchons ainsi des limites de la « langue commune » en étudiant ce qui peut être considéré comme des « langues spécialisées de grande diffusion ».

(8) Cette information, codée sous la forme d'un attribut associé à l'élément SPEAKER (*cf*. Fig. 1), permet d'observer, si elles existent, des corrélations régulières entre le statut des locuteurs et les mots ou expressions employés.

(9) Deux types de prononciations déviantes sont répertoriables : celles qui sont classiquement observées quand on compare un énoncé effectivement produit à ce que serait un énoncé attendu en « français standard », et celles qui sont spécifiques au domaine observé, par exemple ici [teɛfse], qui serait une prononciation classique normale, est une variante de la prononciation footballistique la plus habituelle, semble-t-il, [tefese].

(10) Sont considérées comme étant des « interjections non codifiées », celles qui ne sont pas répertoriées dans l'un des trois dictionnaires généraux monovolumes de référence utilisés, le *Petit Robert*, le *Petit Larousse*, ou le *Dictionnaire Hachette*, ni dans Enckel & Rézeau, 2003.

(11) Pierre Corbin et moi, assistés de Vassil Mostrov, étudiant de la promotion 2002-2003 du DESS LTTAC, puis du DEA « Théories et analyses linguistiques », et maintenant doctorant en linguistique, à l'université Lille III, et vacataire, à l'UMR Silex, dans le cadre du projet OuRAL (campagne Technolangue, sous-projet du projet AGILE).

(12) Ce travail a débuté en 2003-2004.

(13) Je n'entrerai pas ici dans le détail du balisage mis en œuvre sous Transcriber. Les fichiers de transcription intègrent, insérés dans le contenu textuel de chaque élément correspondant à un tour de parole, les éléments COMMENTAIRE proposés par Transcriber, que nous utilisons pour délimiter les informations correspondant aux sous-éléments de l'élément SPEAKER présentés Fig. 1. Ceci permet, pour disposer d'un corpus réunissant toutes les

transcriptions faites à ce jour balisées de manière cohérente et, en attendant l'alignement des fichiers XML du *Corpus foot de multiplex transcrits* avec leurs sources sonores, de rendre compatibles, avec la DTD présentée Fig. 1, les nouvelles transcriptions (celles du *Corpus foot de multiplex transcrits sous Transcriber*).

(14) À ma connaissance, il n'existe pas d'outil disponible offrant l'accès au son à partir d'une concordance compilant les segments textuels extraits d'un corpus XMLisé, comme ceux générés par Transcriber, où sont mentionnées, en attribut, dans les balises de chaque tour de parole, les informations utiles à la synchronisation de la transcription et d'un extrait du fichier son. Si Context, le concordancier développé par Jean Véronis, permet bien d'accéder aux sources sonores à partir des concordances, il n'est pas conçu pour extraire les références des segments sonores alignés du balisage XML des transcriptions : il impose la création d'un troisième fichier décrivant les alignements. Ainsi, avant de pouvoir pleinement évaluer les bénéfices à tirer du *Corpus foot de multiplex transcrits sous Transcriber*, il conviendra d'élaborer une interface de consultation capable de traiter des transcriptions XMLisées alignées à leur source sonore, ou de formater nos données pour les rendres exploitables avec Context.

(15) Les étudiants de la promotion 2002-2003 du DESS LTTAC m'ont donné à observer le travail fait avec et pour Pierre Corbin et/ou moi, et ont ainsi contribué à ma réflexion sur ce point.

(16) On dira qu'*une embarcation remonte un fleuve* quand elle avance dans le sens opposé à celui du courant, en luttant contre son action, que l'*on remonte un quai* quand on marche sur la berge dans le sens opposé à celui du cours d'eau longé, ou que l'*on remonte une rue* quand on gravit sa pente ou, sans incidence de son éventuelle pente, quand on la parcourt dans le sens inverse du flot de voitures, voire des numéros des bâtiments qui la bordent. Les joueurs font de même, ils remontent le terrain contre l'opposition plus ou moins effective des joueurs de l'équipe adverse.

(17) Je ne considère pas ici les questions de coût des logiciels, mais seulement l'investissement en formation que réclament certains outils informatiques.

(18) Il est pour cela possible, soit d'étiqueter en parties du discours les items lexicaux du corpus puis d'utiliser un concordancier qui prenne en compte ces informations, soit d'employer un logiciel comme Intex ou Unitex (développés respectivement par Max Silberztein – dans un premier temps au LADL avec Maurice Gross, puis de manière autonome – et au LADL puis au Laboratoire d'Informatique de l'Institut d'électronique et d'informatique Gaspard-Monge à l'université de Marne-la-Vallée, notamment par Éric Laporte et Sébastien Paumier), qui propage, si on le demande, les informations morphosyntaxiques contenues dans leurs dictionnaires (les DELA) et les codages, en particulier « sémantiques », déclarés dans des dictionnaires personnels ajoutés aux ressources DELA, sur les mots des documents qui sont explorés avec eux (Silberztein *& al.*, 2001). Ce point mérite un développement plus conséquent, qui permette d'évaluer les qualités respectives de ces procédures et des outils à mettre en œuvre dans chaque cas. Il n'a pas sa place dans cette note, mais a déjà donné matière à une publication plus technique (Gasiglia 2004, p. 53-60).

(19) Je reviendrai, dans la section suivante, sur l'importance qu'il y a à ce que les explorations automatiques soient menées avec une curiosité affûtée, et que les extractions produites soient dépouillées en faisant preuve d'une sensibilité linguistique qui, exercée sur ce qui est donné à voir, conduise à envisager de nouvelles requêtes au fur et à mesure que des faits semblent s'observer, afin de les valider ou pas. S'il est vrai que seul peut être trouvé ce qui a été cherché, les motifs de recherche eux-mêmes demandent à être trouvés.

(20) Le rendement de consultation est amélioré si l'on balise les noms de joueurs en mentionnant en attribut l'équipe pour laquelle ils jouent, au moment où est enregistrée la retransmission radiophonique de chaque match. Cette information n'avait pas été balisée dans le premier état du *Corpus foot de multiplex transcrits* (celui exploité en 2002-2003 et dont la

DTD est présentée Fig. 1). Les dépouillements réalisés alors par des personnes néophytes en football ont montré son importance et nous ont incités à envisager de l'introduire dans le balisage. Leur importance s'est révélée encore plus évidente quand, un an plus tard, nous avons décidé d'augmenter le nombre de retransmissions transcrites, que certains joueurs avaient changé d'équipe et que les risques de mauvaise interprétation ont augmenté encore.

(21) Pierre Corbin (à paraître : § 8) reproduit la description de l'emploi transitif du verbe *décaler* proposée par Gaël Gauvin (étudiant du DESS LTTAC promotion 2002-2003), un exemple d'exploitation particulièrement intéressante des occurrences discursives rendues disponibles par le corpus.

(22) Je ne chercherai pas à faire la part des choses entre deux terminologies concurrentes, l'une réunissant sous « collocation » toutes les cooccurrences régulières, l'autre ne retenant que les séquences intégrant des figures, et dont le sens n'est pas déductible de la réunion des sens des différents constituants impliqués. Ici les collocations sont à entendre comme relevant du premier type, ce sont des cooccurrences.

(23) Certains items lexicaux considérés comme « mineurs » (les déterminants et prépositions en particulier) ne seront pas nécessairement non plus pris en compte lors de la recherche de collocateurs : les « collocations » repérées peuvent ainsi correspondre à des constructions différentes.

(24) Il est probable que ces outils aient été conçus pour travailler sur des corpus particulièrement volumineux et que, lorsqu'ils les explorent, la quantité finisse par lisser ces dysfonctionnements. Il n'en demeure pas moins que ces outils traitent la segmentation des énoncés de manière linguistiquement peu pertinente, et qu'exploités sur des corpus de taille modeste ils génèrent du bruit.

(25) L'ontologie des actions de jeu du football a été établie avec la très précieuse collaboration de François et Pierre Corbin, qui, par leur connaissance des règles et stratégies de ce jeu collectif, lui ont donné sa forme première et sa qualité descriptive. Elle a été ensuite interactivement enrichie et corrigée à partir d'observations formulées lors de dépouillements du corpus menées par les étudiants de la promotion 2002-2003 du DESS LTTAC, Pierre Corbin et moi-même. Pierre Corbin (à paraître : § 5) revient sur sa description et son exploitation.

(26) Certains de ces paramètres sont applicables plus largement à tous les sports collectifs de ballon.

(27) Certains de ces mots-clés, comme ceux mentionnés ici, ont une certaine évidence, d'autres, comme *axe* (*dans l'axe*) ou *retrait* (*en retrait*), ne nous apparaissent comme intéressants qu'à condition d'avoir un minimum de familiarité avec le domaine, ce qui pourra difficilement s'obtenir sans avoir fait l'effort, pour le football, de suivre quelques matchs, voire de connaître les règles du jeu. Le corpus, si riche soit-il, ne peut pas se substituer complètement à la compétence du lexicographe qui le consulte.

(28) Le verbe *tacler* observera la même partition sémantique et aura comme équivalents, pour le second type d'emplois, les verbes *sécher* (bien que ce sens ne soit pas relevé dans le *Petit Robert* électronique), *faucher*, etc.

(29) Les segments notés entre parenthèses et séparés par le signe « + » sont en relation de disjonction, l'un d'eux seulement occupe effectivement cette place syntaxique dans les énoncés attestés. « *Ø* » note la possible absence d'un argument en une position donnée : *dégager (le ballon + la balle + Ø)*, équivaut à : *dégager le ballon + dégager la balle + dégager*. Des différences de modalisation sont observables mais non décrites ici : *N-DEFENSE-EQUIPE$_i$ va pouvoir se dégager* ou *N-GARDIEN-DE-BUT-EQUIPE$_i$ (peut dégager + va pouvoir dégager)*.

LE DISCOURS DIRECT DANS LE CORPUS COMPARABLE SPÉCIALISÉ

Aurelija Leonavičienė
Universitas Vytauti Magni

1. INTRODUCTION

La mise en scène de la parole d'autrui est une stratégie discursive fréquente dans l'écriture de la presse actuelle. Les voix-témoins insérées dans le discours monologal du journaliste se présentent comme des médiateurs objectifs transformant l'article en texte à plusieurs voix.

La problématique d'immixtion de voix multiples dans le texte monologal, examinée par M. Bakhtine dans l'analyse des œuvres littéraires (par exemple, de Dostoïevski) dans le but de mieux les caractériser, a été exploitée par beaucoup d'autres linguistes : Julia Kristeva (analyse de l'intertextualité), O. Ducrot (analyse de la polyphonie), Jaqueline Authier (analyse de l'hétérogénéité montrée et l'hétérogénéité constitutive), etc.

L'hétérogénéité des textes se manifeste par la présence de plusieurs voix : l'abondance de citations, des discours rapportés, etc.

Dans cet article, nous nous intéressons à une forme explicite du discours rapporté, codifiée par la grammaire classique sous le nom du *discours direct* (DD). Notre analyse a été largement inspirée par les travaux de J. Authier, L. Rosier, D. Maingueneau et d'autres linguistes.

Le discours direct qui fait l'objet de notre analyse est étudié dans le corpus comparable spécialisé d'une taille d'un million d'occurrences, constitué d'articles de la presse politique française (*Le monde, Libération*) et lituanienne (*Lietuvos rytas*, *Lietuvos žinios*) qui datent de 2001-2003. Ce corpus ne contient ni interviews, ni débats, ni entretiens. Il est constitué d'articles politiques de la presse écrite des genres *information* et *commentaire*.

L'article a pour but de décrire les manifestations du discours direct dans la presse politique française et lituanienne et ses modes de l'oralisation de l'écrit. Dans ce but, nous présentons la définition du DD, analysé à des niveaux différents linguistiques (syntaxe, lexique) qui prouvent la présence du langage oral dans l'écrit.

Le problème de l'oralisation du discours direct dans la presse politique est analysé par le moyen des méthodes descriptive et comparative. L'étude

comparative qui situe nos recherches dans l'axe de l'espace (presse de différents pays) permet de faire apparaître les constantes du sous-genre ainsi que les variantes propres à la presse de chaque pays. Pendant la dernière décennie, après 1991, la presse politique lituanienne a subi beaucoup de changements. Les recherches linguistiques contrastives nous permettent de remarquer certaines tendances de la stratégie discursive de la presse actuelle, lituanienne et française.

Étant véhicule d'information, la presse utilise plusieurs moyens possibles pour introduire des voix-témoins afin de parvenir à son tout premier but, celui d'informer et intéresser. L'analyse du corpus comparable spécialisé montre la présence massive des citations, qui confirment les paroles du journaliste, créent l'effet d'objectivité, d'authenticité. Le discours direct peut aussi être employé pour de pareilles raisons, mais il est utilisé également dans le but de naturaliser les informations racontées en les transposant au registre familier, en oralisant l'écrit et en même temps dans le but de réduire la distance sociale. L'écrit n'est pas une simple représentation de l'oral mais il offre une certaine mise en scène de la parole. Les combinaisons sont nombreuses et nuancent la dichotomie rigide oral/écrit (Rosier, 2000).

Dans les parties suivantes du travail, nous présentons la définition du discours direct, analysons ses formes et les différentes manifestations de l'oralité dans le DD de la presse politique.

2. LE DISCOURS DIRECT, LE DIALOGUE ET LE PSEUDO-DIALOGUE

Le DD, étant une des formes du DR, est défini comme le rapporté, dans sa forme originale, sans termes de liaisons, après un verbe de parole (et, dans la langue écrite, placé entre guillemets) (*Le petit Robert 1*, 1978), ou comme la restitution des paroles exactes de l'énonciateur cité (Maingueneau 2000 ; DLKŽ 2000) marquées typographiquement dans le texte (Labutis, 2002). En fait, le discours direct ne rapporte pas nécessairement des paroles effectivement dites, il peut s'agir d'une énonciation rêvée, future, prescrite, etc., mais ces dernières ne sont pas répérables dans des textes de la presse. Le DD se caractérise en effet par le fait qu'il est supposé indiquer *les mots mêmes* de l'énonciateur cité. Le DD est interprété parfois comme une *théâtralisation*, la mise en scène de la parole d'autrui. Il lui faut donc un espace spatio-temporel « présent », un moi-ici-maintenant sans cesse renouvelé (Rosier, 1996). L'appropriation active théâtrale de la parole d'autrui dans la presse politique fait apparaître l'oral dans l'écrit.

Le discours direct interrompt le discours citant (discours du journaliste) et crée dans celui-ci un autre espace d'énonciation plus ou moins délimité. Les deux voix, celle du journaliste et celle de la voix-témoin, sont nettement distinguées, du moins lorsque le discours direct est nettement distingué du

texte. Ses principaux marqueurs sont le détachement par les deux points et les guillemets, les tirets en cas d'échange dialogué, et généralement une lexie – verbe de communication le plus souvent – qui l'introduit ou le signifie lorsqu'elle est placée en incise (« dit-il »).

La forme classique la mieux connue du discours direct est le *dialogue*. En présentant sous cette forme les paroles d'autrui, l'auteur (le journaliste) prend ses distances, s'il ne participe pas lui-même dans le dialogue, et crée l'effet de conversation. En plus, la forme du dialogue (les tirets, les majuscules, etc.) autonomise le DD, le distingue strictement des autres paroles et conversationnalise le texte écrit. Il le rapproche du code oral.

(1) Pour la rentrée, elle a invité un professeur d'histoire de Washburn à faire un cours sur l'Irak. Il y avait une dizaine d'étudiantes et un étudiant. Carey avait apporté des fraises et des brownies. Les étudiantes étaient perplexes.

« J'étais contre la guerre mais je ne veux pas être antiaméricaine.
- Ça ne devrait pas être antiaméricain de contester son gouvernement. L'Amérique nous offre les libertés, pourquoi pas celle de la critiquer ?
- Moi, ça ne me gêne pas d'être antiaméricaine. Je suis déçue. Je voulais m'engager dans les Peace Corps − coopération. *L'idée que je représenterais l'Amérique me fait honte.*
- Je ne suis pas d'accord. La honte, c'était d'insulter les soldats au retour du Vietnam.
- Est-ce qu'on peut faire la loi dans d'autres pays ? Est-ce que le 11 septembre n'est pas une réponse à ce qu'on a fait au reste du monde ? » (*Le Monde*, 27.03.03)

(2) Po kiekvienos repeticijos specialus asmuo tikrindavo rezultatus − skaičiuodavo, kiek manekenų kliudyta, kur pataikiusios kulkos. Po to « *teroristės* » būdavo perkeliamos į kitas vietas salėje ir viskas kartojama iš naujo.

« *Iš kur žinote, kur sėdės šios moterys ? − paklausiau savo vadovo. − Ir kodėl jos turėtų laukti, kol jūsų vaikinai pasieks jas iš vestibiulio ir koridorių ?* »

« *Daug žinosi − prastai miegosi* », − atsikirto jis. Tačiau netrukus paaiškino : « *Kai ką mes jau žinome, žiūrovų salėje šiandien buvo derybininkai, atkreipė dėmesį į daug ką, paskui mums papasakojo. Be to, vestibiulį šturmuos visai kita grupė. Mūsiškiai pasirodys... (jis šyptelėjo) kone iš po žemių* ». (*Lietuvos rytas*, 28.10.02)

L'analyse du corpus comparable spécialisé d'une taille de presque d'un million d'occurrences n'offre que quelques exemples du discours direct

présenté sous la forme de dialogue classique (dans des textes lituaniens introduit par les guillemets). Ces exemples demandent des commentaires puisqu'il faut expliquer pourquoi le DD en dialogue est assez rare dans la presse politique française et lituanienne. Le dialogue, qui appartient à l'oral, dans ce type de texte représente un écart par rapport à la norme stylistique journalistique. Donc, étant marqué dans la presse politique des genres de l'*information* et du *commentaire*, le DD dialogique est de fréquence assez basse.

Une forme de la présence du DD dans la presse se manifeste encore dans son introduction dans le texte sous la forme de la ***citation***. La plupart des cas du DD trouvés prenant la forme des citations sont introduits par des verbes (le plus souvent par des verbes de communication, centrés sur le message) et des groupes prépositionnels marqués typographiquement. En cas d'absence d'introducteur explicite, les marqueurs typographiques (les deux points et les guillemets) sont les seuls signes qui montrent les frontières du DD dans le texte.

Pour démontrer le dialogisme dans la presse politique, le journaliste regroupe les paroles rapportées (DD) en dialogue. Une telle mise en texte du DD reproduit une conversation ordinaire. Le journaliste se met en scène comme l'interlocuteur ou bien il prend ses distances et laisse parler les autres (crée un dialogue (ou un pseudo-dialogue) de deux ou de plusieurs sources). Les paroles des voix-témoins (DD) regroupées en dialogue sous la forme de citations manifestent l'oralité et la pratique orale (citer des dialogues à l'oral au cours d'un récit est une stratégie narrative très courante). Cette forme de présentation du DD peut être appelée la *simulation d'une interaction* (Tuomarla, 1999), car il ne s'agit pas d'un dialogue oral réel entre les porte-parole. Les citations représentant le DD simulent des échanges conversationnels.

> (3) De hauts fonctionnaires américains ont estimé que la voix de l'homme était « *probablement* » celle de Ben Laden, sans que l'authentification soit certaine. « *C'est probablement lui mais nous ne le savons pas avec certitude* », a dit l'un d'eux. « *On dirait Oussama Ben Laden mais nous ne sommes pas parvenus à cette conclusion* », a dit un autre haut fonctionnaire. (*Le Monde*, 13.11.02)

> (4) Wahid montre la cuisine. Dans un coin, de la bouse séchée. « *Ils mélangeaient à notre nourriture des excréments et de l'urine de chien* », souffle Wahid. Il n'a plus d'ongles. Comment a-t-il tenu ? « *Grâce à Allah. 86 frères ont été raflés avec moi. On est 55 à en être sortis. C'est bien, non ?* » (*Libération*, 11.04.03)

(5) Lietuvos politikai mano, jog neigiamas Airijos referendumo rezultatas politiškai turėtų labai rimtas pasekmes, tačiau nesukurtų « *visiškos aklavietės* ». « *Tai yra rimtas neaiškumas plėtros kelyje, kurį reikia pašalinti* », − sakė Lietuvos vyriausiasis euroderybininkas Petras Auštrevičius. « *Neabejoju, kad plėtrai tai nesutrukdys, bet neigiamas referendumo rezultatas jos ir nepagreitins* », − teigė Užsienio reikalų ministras Antanas Valionis. (*Lietuvos žinios*, 19.10.02)

(6) « *Lietuva tuoj duos pažadą ištekėti, – vakar vakare pasibaigus galutinėms mūsų šalies deryboms su Europos Sąjunga juokavo Lietuvos vyriausiasis euroderybininkas Petras Auštrevičius. – Jaunikis, atrodo, atneša rimtą kraitį. Santuoka bus įpareigojanti, bet žadanti įdomų ir prasmingą gyvenimą* ». *« Esame nuvargę, bet patenkinti. Derybos baigtos sėkmingai. Esame tarp trijų šalių, gausiančių daugiausia lėšų vienam gyventojui* », – kiek santūriau, tačiau taip pat džiaugsmingai kalbėjo diplomatijos vadovas A.Valionis. (*Lietuvos rytas*, 14.12.02)

Les exemples tirés du corpus comparable spécialisé illustrent la simulation des échanges conversationnels et montrent les traits communs aux textes des journaux français et lituaniens. À la lumière des exemples donnés, on remarque que la mise en scène du DD sous forme du pseudo-dialogue peut faire naître la polémique et créer l'impression de communication entre le journaliste et les lecteurs (sur le rapport entre les interlocuteurs).

En plus, l'interaction des voix-témoins (ou de la voix-témoin et ses auditeurs) est renforcée par la création de la situation communicative dans le discours du journaliste. Elle sert à montrer que les paroles du DD introduites dans le texte par les marqueurs typographiques ont été vraiment prononcées, la manière dont elles ont été prononcées et les réactions des interlocuteurs ou des auditeurs suscitées. Le commentaire journalistique crée la situation d'énonciation. Soient des exemples :

(7) « *Turime žengti į priekį, nes kitaip atsiliksime* », − sakė Zeminis. Jo atidžiai klausėsi aukščiausi Kinijos vadovai, įskaitant viceprezidentą Hu Jintao. Šis, kaip manoma, per suvažiavimą turėtų perimti vadovavimą partijai iš pasitraukiančio Zeminio. (*Lietuvos žinios*, 09.11.02, souligné par moi)

(8) Pourtant, le dernier mot revient au spécialiste Brown, chauffeur de camion-citerne : « *Les Etats-Unis, la France, l'Angleterre, l'Allemagne sont les pays les plus riches et les plus puissants du monde. Mais à quoi sert cette puissance si on ne s'en sert pas pour faire le bien, pour libérer les peuples comme les*

Irakiens, qui vivent dans la peur et la misère? » Un silence, puis le spécialiste Brown est ovationné. (*Le Monde*, 17.03.03)

La situation communicative ainsi que le regroupement des citations en dialogue (pseudo-dialogue) créent l'effet de conversation et la polémique dans la presse politique. La présence de l'oral dans le discours direct témoigne aussi le conversationnalisme des textes analysés.

Dans le corpus comparable spécialisé, l'oral (dans le DD) se manifeste le plus souvent à deux niveaux linguistiques : la syntaxe et le lexique.

3. LA MANIFESTATION DE L'ORALITÉ DANS LE DISCOURS DIRECT

3.1. Les particularités syntaxiques, la ponctuation

Un des traits les plus remarquables de l'oralité est la différence syntaxique des phrases du DD par rapport aux phrases du discours citant du journaliste. Le discours direct s'exprime par des phrases courtes, souvent élliptiques, exclamatives, interrogatives dont l'intonation, les pauses et le ton illustrent le caractère spécifique à l'oralité.

(9) Sirven dit qu'il est en train de relire *Un testament espagnol* d'Arthur Koestler, journal d'un condamné à mort pendant la guerre d'Espagne. « *Je ressens les choses comme dans ce livre*, dit-il. *À 76 ans, qu'est-ce qui est important ? L'avenir ? Qui suis-je ? Où vais-je ? Le reste ? Basta ! Vulgaire*. » L'avocat général interrompt la conversation. (*Libération*, 30.01.03)

(10) Devant la Credit Bank of Irak, leur employeur, Ali Faleh et ses deux fils adultes, Hassan et Maitham, observent la scène de loin. Ceux-là refusent de participer à la fête. Membres du Baas ? Fidèles du dictateur ? « *Jamais !* crache Hassan. *Deux de mes oncles ont été exécutés en 1979 par ce fils de p... Qu'il crève !* » (*Le Monde*, 10.04.03)

(11) « *Ir aš šalia jų ? Nacionalbolševikas ? Visiškas absurdas. Provokacija* », − vakar sureagavo A. Kubilius, iš « *Lietuvos ryto" išgirdęs, kur atsidūrė jo pavardė* ». (*Lietuvos rytas*, 24.09.03)

Les exemples de *Libération*, *Le Monde*, *Lietuvos rytas*, *Lietuvos žinios* analysés montrent les aspects syntaxiques mentionnés de l'oralisation du DD. Comme les deux langues analysées diffèrent (le lituanien a le système de cas), alors les formes syntaxiques manifestant l'oral diffèrent aussi. En français la dynamique de l'échange verbal amène à produire des formules phatiques, qui contribuent à maintenir le contact (*bon*, *écoute*, etc.) ou encore

des constructions disloquées à gauche ou à droite, l'extraction (c'est …qui/ c'est… que), certaines phrases sans verbes, etc. Dans le corpus analysé nous avons remarqué l'abondance des constructions syntaxiques mentionnées, surtout des constructions disloquées, si fréquentes à l'oral du français, ainsi que les phrases élliptiques, atypiques et incomplètes propres à l'oral du lituanien.

La construction disloquée est définie comme un syntagme qui se trouve à la droite ou à la gauche d'une proposition et qui est représenté dans cette proposition par un pronom ou un adjectif coréférentiel. Une telle construction est omniprésente à l'oral. Elle permet à l'énonciateur d'indiquer de quoi il va parler, de donner le thème par rapport auquel il organise son énoncé. La fonction syntaxique des éléments détachés est indiquée par des pronoms clitiques de rappel. La construction disloquée est généralement traitée comme un phénomène de la langue parlée ainsi que du style parlé et prête souvent aux paroles citées un ton d'oralité, la dynamique de l'échange verbal.

> (12) La sortie menant à Ramadi, principale ville entre Bagdad et la frontière jordanienne, est contrôlée par un barrage policier. Le responsable ne veut pas dire son nom, mais il insiste pour que ses propos soient rapportés : « *Les Américains mentent, ils disent qu'ils n'en veulent qu'à Saddam Hussein, mais ils bombardent la route. Mais qui l'emprunte, cette route ? Des civils, seulement.* » (*Libération*, 27.03.03, souligné par moi)

> (13) « *Mais c'était à nous, Irakiens, de le renverser. Nous avons essayé et échoué plusieurs fois. Nous aurions dû essayer encore et encore. Maintenant, la victoire est à eux, les infidèles. Pas à nous.* » Hassan a 30 ans, Maithan 25. En d'autres temps, ils auraient pu être ingénieur ou enseignant. (*Le Monde,* 10.04.03)

Dans la langue lituanienne, à l'oral, on remarque une tendance de simplification des constructions syntaxiques ainsi que l'emploi des phrases élliptiques sous la loi de l'économie linguistique. En plus, des phrases atypiques (*benariai sakiniai* – une sorte de phrase impersonnelle (*O ne !/ Kur tau !/ Ne, tikrai ne !* et autres)), ainsi que des phrases incomplètes (*nepilnieji sakiniai*), sont très fréquentes dans le langage parlé. Ces dernières n'ont pas de modèles caractéristiques pour les phrases habituelles. Par conséquent, leur sens dépend de la situation énonciative ainsi que du contexte (Labutis 2002 ; Sirtautas 2001). Les phrases des types mentionnés, introduites dans le discours direct de la presse politique constituent un écart par rapport à la norme syntaxique journalistique, signalent l'oralité ainsi que le ton oral.

(14) « *Lietuviai ? Žinau, prieš kelerius metus mes jus įveikėme futbolo varžybose, o šįkart priimsime į Europos Sąjungą* », − tokiais žodžiais į « *Lietuvos ryto* » korespondentus po šeštadienį Airijoje vykusio referendumo dėl Nicos sutarties kreipėsi į viešbutį vežęs Dublino taksistas. (*Lietuvos rytas*, 21.10.02)

(15) Kancleris pripažino, kad tai, jog šįkart jo partija gavo mažiau balsų negu 1998 metais, kai socialdemokratai užsitikrino 40,9 procento rinkėjų paramą, yra labai skausminga. « *Suprantama, atsakomybė už tai tenka man, − sakė kancleris. − Jei ne man, tai kam ?* » (*Lietuvos žinios*, 24.09.02)

L'intonation et les pauses sont des phénomènes vocaux dont la présence dans l'écrit s'inscrit par les points d'interrogation, d'exclamation et de suspension (ex. 9, 10, 11, etc.). En analysant le corpus comparable spécialisé de la presse, nous constatons que la ponctuation joue un rôle important en ce qu'elle nous permet de visualiser le rythme de l'énoncé cité et, ainsi, montrer comment il est prononcé. Dans des textes écrits, ce langage oral est limité visuellement par des guillemets qui signalent les frontières du discours direct inséré. Alors, en cas d'oralisation du DD dans la presse, les guillemets introduisent une autre intonation, le mimésis verbalisé, qui n'est pas propre au discours du journaliste.

Dans les cas de tous les exemples que nous avons analysés, la ponctuation représente d'un côté le moyen de marquage typographique et de l'autre, le moyen de l'adaptation de l'oral à l'écrit. Le rapport des paroles d'autrui par le DD garde la fidélité à la voix, sa transparence sémantique et vocale. Ici, il ne s'agit pas de la verbalisation du phénomène mais d'une résolution typographique d'un problème posé par rapport à l'oralité : l'usage des capitales, parfois soulignées de gras.

> (16) Il y a Ahmed, qui sort d'une enveloppe les pièces d'identité de ses trois fils, tous mineurs, et dont la voix cassée arrive à peine à hurler. « *On m'a dit qu'on les avait embarqués ici. Où sont-ils ? OÙ SONT-ILS ?* » (*Libération*, 05.04.03)

L'usage des capitales dans le DD, décrit par Rosier (1997), est appelé la *vocifération* – littéralement *transport de voix* – la volonté de rendre volume sonore et tonalité en les matérialisant par le recours à la typographie. Dans le corpus d'une taille de presque d'un million de mots nous n'avons trouvé que des exemples français (de *Libération*).

Pour déterminer la fréquence du discours direct oralisé nous avons étudié une partie du corpus des quatre journaux (de taille similaire pour chacun). Les résultats statistiques de la manifestation de l'oralité dans le discours direct aux niveaux syntaxique et lexical du corpus comparable spécialisé sont présentés dans le *tableau 1*. Cette analyse a été basée sur l'examen des articles écrits sur les mêmes thèmes et problèmes, constituant environ 50 000

occurrences de chaque journal. Comme la longueur des textes des journaux analysés n'est pas la même (les articles du *Monde* sont les plus longs, et ceux de *Lietuvos žinios* les plus courts) alors, nous avons examiné 100 articles de *Lietuvos žinios*, 93 articles de *Lietuvos rytas* ainsi que 75 articles de *Libération* et 53 articles du *Monde*.

Selon le nombre (qui n'est pas très grand) des cas trouvés de l'oralisation du DD au niveau du lexique et de la syntaxe, nous pouvons constater que l'oralité n'est pas très souvent présente dans la presse politique. Les particularités linguistiques du code oral se manifestent plus fréquemment dans les journaux français (*Le Monde*, *Libération*) que dans les journaux lituaniens, où les journalistes préfèrent le DD régi par les normes du code écrit ou l'intégration des paroles d'autrui dans leurs propos sous la forme du discours indirect (surtout dans *Lietuvos žinios*). Étant sérieux et objectif, *Le Monde* pratique un jeu subtil avec les attentes du lecteur (Maingueneau, 2000) et, par conséquent, le code oral ne s'exprime pas par des formes aussi variées et nombreuses que dans *Libération*. Le discours direct de ce dernier journal se présente sous une grande diversité des formes (transport de voix, etc.) qui témoignent du mélange des registres de langue dans le discours journalistique. Par cela, le journaliste montre un ethos d'individu libéré de tout univers verbal clos, de tout préjugé, et crée ainsi un style particulier, appelé un *style Libé* (Maingueneau, 2000).

Nr.	Titre du journal	Nombre d'articles	Nombre d'occurences d'articles analysés	Oralisation du discours direct	
				nombre d'exemples syntaxiques trouvés	nombre d'exemples lexicaux trouvés
1.	***Lietuvos žinios***	100	50215	10	4
2.	***Lietuvos rytas***	93	50222	18	3
3.	***Libération***	75	50253	35	8
4.	***Le Monde***	53	50077	19	7

Tableau 1 : Manifestation de l'oralité dans le DD du corpus comparable spécialisé.

3.2. Le niveau lexical

Comme nous avons pu constater, l'oralisation du DD dans la presse politique se manifeste aussi au niveau du lexique. Les mots empruntés à l'oral (mots familiers, argotiques, etc.), constituent un écart par rapport à la norme du style journalistique des textes politiques. Ils représentent un autre entourage lexical, et signalent généralement la présence de l'oralité. Le

corpus analysé fournit des exemples de la manifestation de l'oralité au niveau du lexique du DD existant dans la presse lituanienne et française.

> (17) Face aux hordes de journalistes qui l'assaillent à l'ONU depuis des semaines, Hans Blix reste imperturbable. Il répond toujours aux questions un demi-sourire aux lèvres, poliment et posément, sans hausser le ton. L'heure a beau être grave, lui est serein. « *C'est peut-être sa principale qualité*, estime un diplomate qui le connaît bien. *Il sait garder son sang-froid quelle que soit la situation. C'est quelqu'un de cool. Remarquez, avec le boulot qu'il a, il vaut mieux. Sinon, on irait à la catastrophe.* » (*Libération*, 28.01.03, souligné par moi)
>
> (18) Samedi, en regardant les « nouvelles » à la télévision, il s'est levé d'un coup puis a fouillé fiévreusement dans l'armoire. Il cherchait sa carte d'électeur. C'est qu'il n'avait pas voté depuis un bail, Aldo. Et dimanche matin, il a été un des premiers à déposer son bulletin. « *Il faut pas le laisser passer, le couillon* », disait-il. (*Le Monde*, 06.05.02)
>
> (19) À 7 heures, un sexagénaire déboule : « *Putain, ces cons, ils vont nous empêcher d'aller manifester !* » Le 7 h 41 vide la salle d'un coup. (*Libération,* 14.05.03)
>
> (20) « *Per šturmą dėl pasklidusių dujų beveik nieko nenutuokėme, daugelis prarado sąmonę. Tik tie įkaitai, kurie sugebėjo kuo nors užsidengti nosį ir burną, dar kažką girdėjo ir jautė. Kiti iš karto "atsijungė"* », − sakė jaunuolis. (*Lietuvos rytas*, 28.10.02)
>
> (21) Tačiau Brazauskas sakė : « *Jokių intrigų mes nepiname. Dar kartą galiu pasakyti : Grigaravičius kaip generalinis komisaras dirbo gerai. Paprastai šnekant, kai kas jį "išdūrė", ir "išdūrė" jį politikai.* » Tačiau jis atsisakė atskleisti, apie kokius politikus kalbama. (*Lietuvos žinios*, 11.04.03)

Les mots familiers (*boulot*, *couillon*, *cool*, *etc.*), populaires et argotiques (*išdūrė*, *atsijungė etc.*) trouvés dans le corpus analysé montrent la présence du style familier, ou du style populaire, dans le discours direct des textes de la presse politique. L'introduction du DD manifestant l'oralité dans le style journalistique est une stratégie discursive de la presse actuelle. L'observation de la presse politique lituanienne et française permet de constater que l'insertion d'un style dans un autre style donne la possibilité au journaliste d'établir la distance avec les individus parlants, et de restituer leur point de vue par les mots prononcés. De plus, les mots du DD guillemeté appartenant au style familier ou populaire jouent le rôle de provocateur dans la presse politique officielle. Le lexique, les déclarations ou généralement les paroles

d'autrui qui soulignent le conflit provoquent, à leur tour, le conflit (Lorda, 2000). La mise en relief des mots ou des propos choquants et agressifs suscite immédiatement des réactions, et contribue à créer de nouvelles polémiques. L'exemple lituanien ici présenté *išdūrė* (21) peut être considéré comme l'illustration de l'élément provocateur lexical. Introduit par le journaliste entre les guillemets, pour ne pas trop choquer le lecteur, et appartenant à l'oral argotique, ce mot, ainsi que les phrases environnantes, prononcées par le premier ministre lituanien, ont suscité une réaction importante des lecteurs et la polémique dans les médias (la télévision, la presse) au mois d'avril 2003.

Nos recherches du corpus comparable spécialisé montrent que les voix des personnes au pouvoir sont parfois représentées dans la presse dans une forme de langue de tous les jours. Selon Fairclough (1995), ce mélange des pratiques du domaine privé avec celles du domaine public naturalise les informations racontées, conversationnalise le texte et réduit la distance sociale entre le locuteur et des lecteurs. La représentation des paroles des politiciens en registre familier comprend souvent des éléments particuliers à l'oral. Le choix et l'utilisation des mots concrets trahissent le caractère individuel de l'acte de parole de l'énonciateur, et se réfèrent à la société d'origine. En comparant les paroles des voix-témoins (DD) introduites dans la presse politique, on arrive à la conclusion que les paroles rapportées de *vox populi* reflètent l'oral plus souvent que les paroles des hommes politiques. Le choix de la source citée est considéré comme une démarche importante pour mettre en valeur la nature orale de l'énoncé cité.

Le ton oral peut être prêté au discours direct en reproduisant des particules énonciatives ou des mots du discours qui contribuent à créer une image dialogique de la situation d'énonciation originale (Tuomarla, 1999). Les particules énonciatives portant le caractère émotionnel et expressif n'ont pas le plus souvent de signification nominative et, par conséquent, appartiennent à la périphérie des mots (Paulauskienė, 1994). Précédant d'autres mots (les porteurs de sens), les particules énonciatives sont étroitement liées à l'intonation, aux gestes, aux mimiques. Elles sont couramment utilisées dans le langage parlé et exprimées, morphologiquement, par les interjections, les onomatopées, etc.

Dans notre étude nous privilégions l'interjection qui est relativement plus fréquente que les autres particules énonciatives dans le DD du corpus analysé. Nous la définissons selon les critères des travaux linguistiques actuels, qui soulignent le plus souvent deux traits communs : son naturel et son expressivité. L'interjection paraît être liée à la mise en voix des affects du sujet, et est donc perçue comme une verbalisation « spontanée ». Celle-ci se trouve sous une forme ponctuée à l'écrit où l'interjection devient essentiellement exclamative (Rosier, 2000). Selon Barbéris (1995), à l'oral, l'interjection n'est pas toujours exclamative, mais elle peut jouer un rôle de

régulateur dans la conversation et perdre le rôle expressif qui lui est propre. Nos recherches et l'analyse des exemples du corpus montrent plutôt son caractère expressif et émotionnel.

> (22) Dans le New Hampshire, par exemple, un Etat qualifié de décisif, John Sununu, fils de l'ancien secrétaire général de la Maison Blanche sous George Bush père, a réussi à conserver un siège républicain convoité par les démocrates. En Caroline-du-Nord, Elizabeth Dole, femme de l'ex-sénateur et ancien candidat à la présidence Bob Dole, s'est imposée. « *Ouaah, quelle nuit !* », a-t-elle lancé après sa victoire. (*Libération*, 07.11.02)

> (23) À la question de savoir si ce résultat pourrait être atteint sans qu'une guerre soit déclenchée contre Saddam Hussein, M. Rumsfeld a répondu ironiquement. « *Oh ! oui, bien sûr,* a-t-il dit, *Saddam Hussein pourrait décider que son avenir est limité et qu'il a envie de partir*. » (*Le Monde*, 19.09.02)

Nous traitons l'interjection dans le cadre de la représentation de l'oral dans l'écrit. Les exemples fournis par le corpus comparable spécialisé montrent son aspect émotionnel. La présence de l'interjection exprimant les émotions signale l'existence de la situation d'interlocution. Généralement, le DD est utilisé dans la presse pour reproduire les paroles réelles qui n'ont pas le caractère fictif propre aux textes littéraires. Par conséquent, l'interjection y joue un rôle important, en mimant la spontanéité du locuteur qui énonce, et en souligne la relation avec l'oral. L'emploi de l'interjection dans la presse écrite ne se limite pas à la manifestation de la spontanéité ou de l'expressivité, mais on y connaît un certain taux d'informativité. L'analyse des exemples de la presse nous permet de constater que, dans le discours direct, l'interjection exprime la subjectivité à visée argumentative, mais elle peut aussi être utilisée comme un morphème désémantisé et purement intonatif, propre à l'oral. En plus, l'interjection authentifie l'acte de parole et sert de garant de la vérité informative ou, selon Rosier (2000), elle est mentionnée pour faire vrai, pour montrer qu'une parole existe, a été prononcée. Le nombre d'exemples trouvés dans le corpus mène à des réflexions sur la fréquence basse d'interjections ainsi que d'autres particules énonciatives, des mots qui prêtent un ton oral au discours direct de la presse politique.

4. CONCLUSION

L'analyse de l'introduction du DD dans la presse politique actuelle, ainsi que de l'oralité transcrite à l'intérieur du discours direct, montre l'influence de l'oral sur l'écrit. En authentifiant les paroles d'autrui, le journaliste cherche le compromis entre deux codes : l'oral et l'écrit. Cette stratégie se

manifeste dans le discours citant, puisque tantôt le journaliste garde la position oficielle, tantôt il mime et adapte son discours au DD exprimant l'oralité. Les commentaires journalistiques rapprochant le discours direct (l'oral) du discours citant (l'écrit) réduisent leur distance, ainsi que la distance sociale entre le locuteur et les lecteurs ordinaires. L'analyse de l'oralité du DD dans le corpus comparable spécialisé permet de constater une certaine neutralisation de l'opposition oral/écrit dans la presse politique française et lituanienne, surtout dans le cas des paroles de *vox populi*.

L'étude du DD dans le cadre de l'oralité transcrite montre la stratégie discursive de la presse politique française et lituanienne. Le discours direct est un garant de l'objectivité, de l'image du vérisme, mais parfois il présente la situation virtuelle, le rapport des paroles individualisées, émotionnelles (la présence d'interjections), et donc reflète le point de vue subjectif. Notre étude des niveaux différents (syntaxe, lexique) où se manifeste l'oralité dans le discours direct nous permet de constater la présence de subjectivité, ainsi que des cas de multiplicité d'opinions du discours de la presse. Le DD présenté sous la forme des citations en dialogue manifeste également un certain « dialogisme », et rend le texte polémique.

Les exemples tirés du corpus comparable spécialisé illustrent les traits communs aux textes politiques lituaniens et français : l'hétérogénéité exprimée par la présence du DD, la manifestation des paroles d'autrui (DD) sous la forme du dialogue, ou de citations simulant parfois des échanges conversationnels, la création de la situation communicative dans le discours du journaliste etc.

L'analyse de l'oralité du DD montre les stratégies discursives de la presse politique des quatre journaux étudiés et fait apparaître des variantes, propres à chaque langue, aux niveaux de la syntaxe et du lexique. Selon la fréquence et la diversité des formes manifestant l'oralité, nous pouvons arriver à la conclusion que *Libération* exprime mieux l'oral que les autres journaux analysés et représente l'ethos de l'individu libéré. À la différence de la presse française, où on peut trouver plusieurs articles sur le sujet donné, la presse politique lituanienne présente le plus souvent un seul article par jour sur ce sujet donné (surtout *Lietuvos žinios*), et communique fréquemment la réaction officielle. Et par conséquent, celle-ci manifeste moins l'oralité transcrite dans le discours direct.

Tous les cas trouvés de l'oralité transcrite dans le discours direct expriment un rapport immédiat au vécu, donnent au lecteur l'impression d'être présent dans la situation, et confèrent un caractère oral spontané aux phrases guillemetées et conversationnalisent les textes.

5. RÉFÉRENCES

Authier, J. 1981. « Hétérogénéité montrée et hétérogénéité constitutive : éléments pour une approche de l'autre dans le discours », *DRLAV*, n° 26.

Authier, J. 1981. « Paroles tenues à distance », *Matérialités discursives*. Lille : Presses Universitaires de Lille. p. 127-142.

Bakhtine, M. 1984. *Esthétique de la création verbale*. Paris : Gallimard.

Barbéris, J.M. 1995. « L'interjection : de l'affect à la parade, et retour », *Faits de langues*, n° 6, p. 93-104.

Čermák, F. 1999. *Kalbos tyrimo metodai*. Kaunas : Vytauto Didžiojo universiteto leidykla.

Dabartinės lietuvių kalbos žodynas. Vilnius : Mokslo ir enciklopedijų leidybos institutas. 2000.

Ducrot, O. 1984. « Esquisse d'une théorie polyphonique de l'énonciation », *Le Dire et le dit*. Paris : Minuit. p. 171-233.

Fairclough, N. 1995. *Media Discource*. New York : Edvard Arnold.

Kristeva, J. 1970. *La Poétique de Dostoïevski*. Paris : Éditions du Seuil.

Labutis, V. 2002. *Lietuvių kalbos sintaksė*. Vilnius : Vilniaus universiteto leidykla. p. 137-138, p. 356-364.

Le petit Robert 1. Paris : S.N.L. 1978.

Lorda, C.-U. 1997. « La relation de déclaration politiques : hétérogénéité et mise en scène de la parole », *Pratiques*, n° 94, p. 62-74.

Lorda, C.-U. 2000. « Les articles dits d'information : la relation de déclarations politiques », *Semen*, n° 13, p. 119-134.

Maingueneau, D. 1994. *Syntaxe du français*. Paris : Hachette. p. 39-43, p. 133-135.

Maingueneau, D. 1996. *Les termes clés de l'analyse du discours*. Paris : Éditions du Seuil.

Maingueneau, D. 2000. *Analyser les textes de communication*. Paris : Nathan/HER. p. 115-147.

Marcinkevičienė, R. 2002. « Palyginamieji tekstynai – šaltinis tarptautinių žodžių vartosenai tirti », *Kalbotyra*, n° 51 (3), p. 81-93.

Monville-Burston, M. 1993. « Les *verba dicendi* dans la presse d'information », *Langue française*, n° 98, p. 48-66.

Paulauskienė, A. 1994. *Lietuvių kalbos morfologija*. Vilnius : Mokslo ir enciklopedijų leidykla. p. 416-420.

Rosier, L. 1993. « De la stylistique sociologique suivie d'une application pratique : discours direct, presse et objectivité », *Revue belge de philologie et d'histoire*, n° 71:3, p. 625-644.

Rosier, L. 1996. « Ces mots qui ne vont pas de soi et ces mots qui ne sont pas de soi », *Travaux de linguistique*, n° 32, p. 155-167.

Rosier, L. 1997. « Entre binarité et continuum. Une nouvelle approche théorique du discours rapporté ? », *Modèles linguistiques*, n° 35, XVIII, p. 1, p. 7-16.

Rosier, L. 2000. « Interjection, subjectivité, expressivité et discours rapporté à l'écrit : petits effets d'un petit discours », *Cahiers de praxématique*, n° 34, p. 19-49.

Sirtautas, V. 2001. « Teksto sintaksė ir vientisinio sakinio analizė », *Rinktiniai kalbotyros straipsniai*. Šiaulių universiteto leidykla. p. 20-41.

Tuomarla, U. 1999. « Le discours direct de la presse écrite : Un lieu de l'oralisation de l'écrit », *Faits de langues*, n° 13, p. 219-229.

6. INDEX

Construction disloquée, conversationnalisation, dialogue, discours direct, discours rapporté, entourage lexical, hétérogénéité constitutive, hétérogénéité montrée, particularité énonciative, phrase incomplète, pseudo-dialogue, simulation d'une interaction, situation énonciative, situation communicative, vociférération.

UTILISATION D'UN CORPUS CATÉGORISÉ POUR L'ÉTUDE ET LA REPRÉSENTATION DE LA SYNONYMIE EN CONTEXTE.

Jean-Luc Manguin
Laboratoire CRISCO - Université de Caen

1. INTRODUCTION

Le but de cet article est de montrer qu'un corpus catégorisé comme *Frantext* peut constituer l'un des supports de l'étude d'une relation paradigmatique avec des visées pédagogiques ou comparatives. Le principe de la méthode est de mettre à l'épreuve les données relationnelles fournies par l'autre fondement de la méthode (un dictionnaire spécialisé), en les confrontant avec les cooccurrences repérées dans le corpus, grâce à la catégorisation des unités lexicales et selon un schéma syntaxique précis. Il est évident que cette catégorisation est insuffisante pour détecter certaines fonctions syntaxiques, mais elle permet dans notre cas de capter la fonction d'adjectif épithète qui nous servira d'exemple. Nous montrerons que notre procédé aboutit à des représentations graphiques d'un grand intérêt pédagogique, ainsi qu'à des mesures quantitatives caractérisant certains emplois de l'unité étudiée. L'exemple choisi est l'adjectif *curieux*, qui est suffisamment courant pour fournir un nombre de données autorisant certains tests statistiques, et qui possède un intérêt linguistique dans le changement de sens induit par son changement de position dans certains contextes nominaux ; par exemple, dans le cas de la synonymie que nous avons choisi d'étudier, un *curieux* regard est probablement un regard intéressant, tandis qu'un regard *curieux* sera plutôt un regard intéressé.

2. LES DONNÉES PARADIGMATIQUES

Celles-ci sont issues d'un dictionnaire de synonymes construit au laboratoire CRISCO à partir des données provenant de sept dictionnaires de référence, et déjà mentionné dans d'autres publications (le modèle est décrit dans Ploux et Victorri, 1998) ; l'exploitation des données présentes dans ce dictionnaire se fait grâce aux outils de la théorie des graphes : nous formalisons le

dictionnaire comme un graphe dont les sommets sont constitués par les mots-vedettes du dictionnaire, et dont les arêtes sont les relations existant entre ces mots (pour la terminologie relative aux graphes, voir Berge, 1958). La relation ainsi définie est réflexive, et symétrique ; ce formalisme a été utilisé en premier par Kahlmann (1975). Pour étudier une unité, on se limite au sous-graphe dont l'ensemble des sommets est composé de l'unité étudiée et de ses synonymes, et dont l'ensemble des arêtes est constitué par toutes les relations existant entre ces sommets (la variabilité du sous-graphe se rencontre dans Manguin et Victorri, 1999). Pour l'adjectif *curieux*, le sous-graphe ainsi défini possède l'allure suivante :

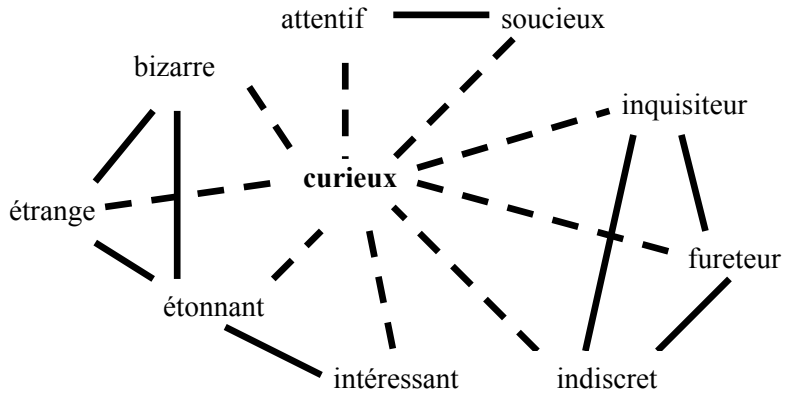


Figure 1 : Structure simplifiée du sous-graphe de synonymie de curieux

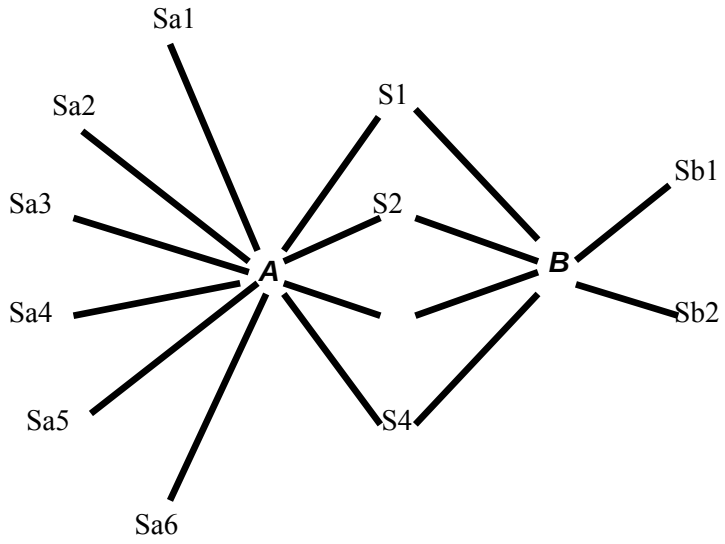
Dans la réalité de notre dictionnaire, *curieux* possède 34 synonymes adjectivaux, que l'on peut, si l'on exclut du graphe les arêtes passant par le mot-vedette étudié, séparer en trois composantes connexes, que nous donnons ci-après :

> 1. amusant, attachant, bizarre, drolatique, drôle, déconcertant, extraordinaire, extravagant, incompréhensible, inconcevable, incroyable, inouï, intéressant, original, paradoxal, piquant, pittoresque, plaisant, rare, singulier, surprenant, unique, étonnant, étrange
> 2. chercheur, fureteur, indiscret, inquisiteur, investigateur
> 3. anxieux, attentif, avide, intéressé, soucieux

Nous voyons que les composantes 2 et 3 correspondent au sens *intéressé* de notre adjectif, tandis que la composante 1 reflète le sens *intéressant*.

Pour construire la représentation graphique à laquelle nous voulons aboutir, nous devons établir une relation valuée entre les synonymes ; Brodda et Karlgren (1969) ont déjà proposé une méthode répondant à cette préoccupation, mais elle est relativement délicate à mettre en œuvre car elle

doit s'appliquer à l'ensemble du graphe formé par le dictionnaire entier. Notre démarche consiste à partir de la matrice d'adjacence du sous-graphe, et pour chaque paire de sommets A et B, nous définissons une similitude entre ces deux sommets par l'indice de Jaccard S (Legendre et Legendre, 1998) :



$$S = \frac{a}{a+b+c} = \frac{4}{4+2+6} = 0{,}333$$

Figure 2 : Exemple de graphe illustrant le calcul de la similitude entre sommets

Dans la formule de calcul, a est le nombre de sommets en relation avec A et avec B, b (resp. c) le nombre de sommets en relation avec A mais pas avec B (resp. avec B mais pas avec A). En effectuant ce calcul pour chaque paire de sommets, nous obtenons une matrice de proximités, à partir de laquelle, par échelonnement multidimensionnel (non-metric multidimensional scaling), nous pouvons construire une représentation bidimensionnelle dans laquelle les synonymes s'agencent en fonction des valeurs des proximités (les calculs sont effectués par le logiciel UCINET 6.0, selon une méthode proposée par Kruskal et Wish, 1978). La représentation de *curieux* et de ses 34 synonymes, que nous donnons en annexe (figure 3), montre ainsi un espace clairement séparé en trois zones :

- l'une contenant les synonymes apparentés à *intéressant*, comme *étonnant*, *étrange*, *paradoxal*, etc.

- une deuxième qui inclut les synonymes comme *intéressé*, *attentif* ou encore *soucieux*.
- une troisième où ce second sens est parfois marqué par une appréciation négative, avec des termes comme *indiscret*, *fureteur*, ou *investigateur*.

Il est tout à fait logique que ces trois zones correspondent aux trois composantes connexes signalées plus haut, puisque l'absence de liaison entre deux sommets qui appartiennent à deux composantes connexes différentes induit une similitude nulle, et par conséquent une distance maximale, entre ces deux sommets. En outre, les deux dernières zones se situent dans une moitié de la représentation, opposées à la première zone, et contiennent des termes synonymes de *curieux* quand celui-ci est en position postposée et s'applique à des substantifs « potentiellement actifs » ; autrement dit, on retrouve sur cette représentation la coupure principale entre les deux sens *intéressé* (dans la partie droite) et *intéressant* (dans la partie gauche).

3. LES DONNÉES SYNTAGMATIQUES

Pour étudier l'adjectif *curieux* en épithète et en contexte nominal, nous faisons appel à un corpus qui nous permet de repérer simplement les groupes Nom + Adjectif ou Adjectif + Nom ; nous avons utilisé la base *Frantext* catégorisée, en excluant le genre poétique, entre 1830 et 1999, ce qui correspond à un corpus de 1 711 textes et 120 millions de mots.

Le recueil des données se déroule en deux étapes. Tout d'abord, nous recherchons dans la base *Frantext* catégorisée les occurrences de *curieux* avec un substantif quelconque, au moyen des requêtes suivantes :

&e(g=A c=&mcurieux) &e(g=S)	(antéposition)
&e(g=S) &?(&e(g=Adv) &e(g=A c=&mcurieux)	(postposition)

On peut remarquer au passage qu'il est nécessaire d'inclure dans la requête des occurrences de *curieux* postposé un adverbe éventuel entre le substantif et l'adjectif. Après rapatriement des 3 500 résultats donnés par *Frantext*, nous effectuons un traitement local par des programmes élaborés au laboratoire par nos soins, et qui permettent de lemmatiser les substantifs rencontrés, afin d'établir la liste des substantifs les plus fréquents employés avec *curieux*, et que nous donnons dans le tableau suivant :

nom	nb. d'occurrences	% sur 3531 occurrences
chose	262	7,42 %
regard	100	2,83 %
fait	66	1,87 %
phénomène	65	1,84 %
spectacle	54	1,53 %
détail	51	1,44 %
œil	45	1,27 %
esprit	42	1,19 %
mélange	39	1,10 %
histoire	35	0,99 %
bête	30	0,85 %
cas	29	0,82 %

Tableau 1 : Les 12 premiers substantifs associés préférentiellement à curieux

Pour notre étude, nous avons en définitive retenu 25 substantifs, qui sont : chose, regard, fait, phénomène, spectacle, détail, œil, esprit, mélange, histoire, bête, cas, exemple, livre, type, expérience, personnage, étude, objet, tête, travail, figure, conversation, spécimen et page. On peut noter au passage que cette liste contient plusieurs noms comme regard, œil, esprit ou bête avec lesquels curieux pourra changer de sens suivant sa position, comme signalé en introduction.

Munis de cette liste de substantifs, et de la liste des synonymes, nous créons tout d'abord deux listes (« adj » et « nom »), puis nous interrogeons à nouveau la base *Frantext*, afin d'obtenir cette fois les cooccurrences nécessaires à notre méthode. Cette fois les requêtes sont de la forme :

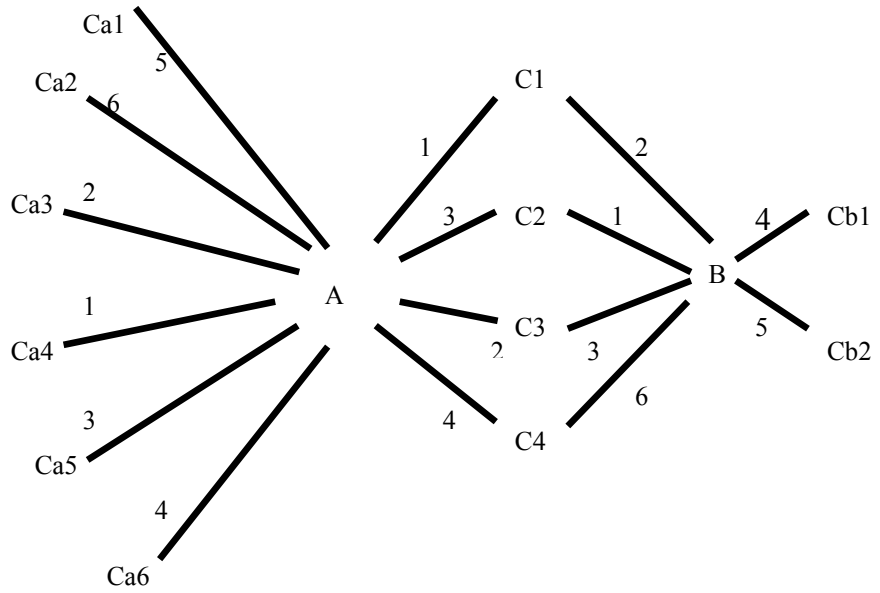
&e(g=A c=&ladj) &e(g=S c=&lnom) (antéposition)
&e(g=S c=&lnom) &?(&e(g=Adv) &e(g=A c=&ladj) (postposition)

Nous recueillons ainsi un total de 6 547 cooccurrences, dont 1 095 concernent *curieux*, et dont le dépouillement automatique selon la même méthode que précédemment nous permet d'aboutir à un tableau (synonymes x substantifs) où chaque case contient le nombre de cooccurrences rencontrées pour la paire envisagée. Nous avons dans notre exemple un

tableau à 35 lignes (*curieux* et ses synonymes) et 25 colonnes (substantifs), dont de nombreuses cases sont vides, puisque sur les 1 750 paires Nom+Adjectif possibles, seules 414 sont effectivement attestées dans notre corpus, ce qui reflète ce que Coseriu nomme les « solidarités lexicales » de type « sélection » (Coseriu, 2001). Ce tableau peut s'assimiler très facilement à la matrice d'adjacence d'un graphe valué, si l'on considère que ce graphe représente la relation de cooccurrence, et que chaque arête porte comme valeur le nombre de cooccurrences pour la paire considérée.

4. LES SIMILITUDES D'EMPLOI

De la même manière que nous avons défini une « similitude paradigmatique », nous pouvons à présent définir une « similitude syntagmatique » ou « similitude d'emploi » entre les synonymes à partir des données du tableau des cooccurrences. Toutefois, comme il s'agit de comparer les préférences d'emploi de chaque adjectif avec chaque substantif de la liste, il faut pondérer les données pour ne pas tenir compte de la fréquence absolue de chacun des adjectifs ; nous ramenons ainsi chaque valeur du tableau à un pourcentage, calculé en divisant cette valeur par le nombre total d'occurrences de l'adjectif. Nous calculons ensuite la similitude d'emploi en utilisant un indice adapté aux données numériques (en l'occurrence l'indice de Steinhaus, *cf*. Legendre et Legendre 1998). Cet indice est basé sur le même principe que celui de Jaccard, c'est-à-dire qu'il divise la partie commune (ici, pour les deux synonymes dont on calcule l'indice, le nombre minimum des cooccurrences observées avec chaque substantif) par la partie totale (le nombre de cooccurrences relevées avec chaque substantif, et l'un ou l'autre des deux synonymes), selon le schéma explicatif ci-contre :



$$S = \frac{2 * Nb\_cooc\_communes}{Nb\_total\_cooc} = \frac{16}{52} = 0{,}31$$

Figure 4 : Exemple de graphe illustrant le calcul de la similitude d'emploi

Pour l'exemple présenté sur la figure 4 (ci-dessous), le calcul de l'indice se fait ainsi :
- au numérateur, on cherche la valeur minimum des deux arêtes partant d'un contexte commun (par exemple pour C4, le minimum est 4), puis on fait le total de ces minima ;
- au dénominateur, on totalise les valeurs des arêtes partant de A ou de B.

À l'issue de cette étape, nous sommes en mesure d'appliquer à notre tableau la même transformation non-linéaire, qui nous conduit à une représentation plane, situant les synonymes les uns par rapport aux autres (voir figure 5 en annexe) ; ce qui importe ici, c'est la confrontation avec les données paradigmatiques. Nous observons en effet que la dissociation entre le domaine *intéressant* et le domaine *intéressé* existe toujours, et qu'aucun terme n'a changé de zone. Par contre, la dichotomie présente au sein de la zone *intéressé* entre les termes dépréciatifs (comme *fureteur*) et les termes neutres (comme *intéressé*) n'est plus présente. Cette différence est probablement due au fait que les données syntagmatiques ne permettent sans doute pas, pour des raisons qu'il importe de préciser, d'accéder à l'essentiel de l'information paradigmatique. Cette conclusion est fondée sur le fait que les regroupements, obtenus par classification hiérarchique ascendante des similitudes d'emploi, présentent parfois des anomalies sémantiques, si on les

compare aux groupes que l'on obtient à partir des similitudes paradigmatiques. Nous pouvons en effet effectuer une classification hiérarchique à partir des tableaux de similitudes, comme nous l'avons déjà présenté par ailleurs (Manguin, 2004). Il est vrai que, dans le cas présent, la seule information paradigmatique incluse dans les données est la relation de synonymie entre *curieux* et ses synonymes (puisque nous avons conservé la liste de ceux-ci), et que les relations entre ces synonymes ont disparu. Des groupes comme (*inquisiteur*, *anxieux*) ou (*avide*, *investigateur*), issus des similitudes d'emploi, nous paraissent en effet moins cohérents que (*avide*, *intéressé*) et (*inquisiteur*, *fureteur*), qui eux proviennent des similitudes paradigmatiques. La solution à ces différences consiste à pondérer les données syntagmatiques par les données paradigmatiques, en faisant simplement, case à case, la moyenne arithmétique des deux tableaux de similitudes ; on obtient alors une représentation plus conforme à la réalité des données de chaque point de vue, qui ressemble beaucoup à celle réalisée avec les seules similitudes d'emploi, mais cette fois les classifications hiérarchiques construites avec ces similitudes moyennes ne font pas apparaître de regroupements aberrants.

D'un autre côté, les valeurs des similitudes d'emploi peuvent révéler certains choix lexicographiques contestables et, par là même, invalider un lien synonymique avec le mot étudié si ces valeurs s'avèrent trop faibles. Par exemple, nous avons éliminé de notre liste de synonymes des adjectifs qui avaient des similitudes d'emploi extrêmement faibles avec *curieux*, comme *sensationnel* (sim = 0,08) et *désireux* (sim = 0,06). Le premier n'est mentionné que par un seul de nos dictionnaires sources (voir la liste de nos dictionnaires dans la bibliographie), tandis que le second correspond à un sens qualifié de « vieux » de notre adjectif *curieux.* En outre, la répartition en deux groupes de synonymes, au lieu de trois, laisse penser qu'une connexité entre tous les termes plus ou moins synonymes d'*intéressé* reflèterait peut-être mieux la réalité du paradigme étudié.

5. LA REPRÉSENTATION DU CONTEXTE

Sur la représentation graphique précédemment construite, et où figurent les synonymes, il est facile de placer les substantifs dont nous avons étudié les cooccurrences. L'emplacement d'un substantif sera le barycentre des points correspondant aux synonymes qui se rencontrent avec ce substantif, affectés chacun de leur nombre de cooccurrences respectif. Il nous faut cependant effectuer une pondération en divisant, pour chaque substantif S, le nombre de ses cooccurrences avec un adjectif par le nombre total d'occurrences de ce substantif avec tous les adjectifs de la liste.

Le placement de ces substantifs fait parfaitement ressortir leurs « classes » (voir figure 6 en annexe) ; en effet, comme nous l'avons déjà dit,

nous avons une représentation où figurent d'un côté les synonymes autour d'*intéressant*, et de l'autre ceux proches d'*intéressé* ; les substantifs se placent eux aussi dans cet espace en deux zones bien séparées

- côté *intéressé* : *œil*, *regard*, *étude*, *bête*, *esprit*, *tête*.
- côté *intéressant* : *cas*, *fait*, *exemple*, *expérience*, *histoire*, *conversation*, *détail* (entre autres).

Pour ne pas surcharger la représentation, nous n'avons pas placé toutes les étiquettes des points correspondant aux substantifs ; par ailleurs, le « groupe des cinq » ne se situe pas exactement dans la zone d'*intéressé*, mais plutôt à mi-chemin d'*intéressé* et d'*intéressant* ; cette position intermédiaire s'interprète comme reflétant le possible changement de sens de *curieux* en position postposée avec ces substantifs (voir le paragraphe 5 qui explique cette conclusion).

Au point de vue lexicologique, il est intéressant d'analyser en détail la position de chacun des termes ; dans le cas d'*étude*, par exemple, le changement de position de *curieux* peut induire non seulement un changement de sens de l'adjectif, mais si c'est le cas, il provoque en plus un changement de sens radical du substantif. En effet, avec *curieux* antéposé, *étude* correspond au produit d'une action intellectuelle, tandis qu'avec *curieux* postposé, *étude* peut désigner l'action intellectuelle elle-même. Pour les autres substantifs de ce groupe, le changement de position de *curieux* n'entraîne qu'une différence de nuance pour le sens du substantif.

6. REMARQUES SUR LA POSITION DE L'ADJECTIF CURIEUX EN ÉPITHÈTE

Comme nous l'avons déjà dit, l'adjectif *curieux* possède, dans certains contextes, la particularité de changer parfois de sens en changeant de position par rapport à son substantif régissant. En séparant les données des cooccurrences de *curieux* en deux groupes suivant sa position, nous pouvons calculer comme précédemment, pour chaque synonyme, les deux indices de similitude d'emploi entre ce synonyme et *curieux* en position anté- ou postposée. La comparaison de ces deux séries d'indices montre principalement que la similitude d'emploi avec *curieux* antéposé est un facteur qui différencie les deux groupes de synonymes présents sur la représentation ; de plus, cette différenciation correspond exactement avec la coupure pseudo-homonymique entre les deux sens de *curieux*, comme le montre le tableau ci-après, où les synonymes sont classés par valeur de similitude (avec *curieux* antéposé) croissante :

Synonyme	**similitude avec *curieux* antéposé**	**similitude avec *curieux* postposé**
Inquisiteur	0,020	0,175
Chercheur	0,020	0,128
investigateur	0,037	0,222
Fureteur	0,037	0,215
Avide	0,053	0,278
Soucieux	0,076	0,253
Intéressé	0,086	0,287
Anxieux	0,087	0,255
Attentif	0,149	0,293
Indiscret	0,155	0,279
inconcevable	0,248	0,394
Drôle	0,263	0,557
Incroyable	0,288	0,476

Tableau 3 : Similitudes d'emploi avec curieux antéposé ou postposé

La coupure entre *indiscret* et *inconcevable* étant fortement significative, il apparaît clairement que les dix premiers adjectifs de ce tableau ne correspondent jamais au sens que recouvre *curieux* antéposé ; comme nous le verrons dans la conclusion, ce fait peut constituer un préambule à un traitement automatique de la désambiguïsation de *curieux*. Mais d'autre part, il faut aussi tenir compte du fait que ces dix adjectifs ont également une similitude assez faible avec *curieux* postposé, et que seuls *attachant* (0,098), *drolatique* (0,225) et *pittoresque* (0,314) ont des valeurs comparables. Cependant, cette apparente anomalie s'explique très bien si l'on se rappelle que *curieux* postposé porte l'un ou l'autre des deux sens possibles, et que parmi les 25 substantifs retenus, 6 seulement sont susceptibles de donner lieu à un changement de sens de *curieux* quand il est postposé ; en termes de cooccurrences relevées de *curieux*, cela ne représente que 24 %, ou si l'on est plus optimiste, 29 % de celles où *curieux* est en position postposée.

7. CONCLUSION ET PERSPECTIVES

Nous avons montré qu'il est possible, en combinant un dictionnaire et un corpus, autrement dit une ressource paradigmatique et une ressource syntagmatique, d'établir une relation valuée entre tous les synonymes d'un même mot-vedette ; en outre, nous avons également montré que cette relation peut aboutir de manière relativement simple à une représentation à but pédagogique. Cette représentation peut de plus s'enrichir de données contextuelles qui contribuent à en améliorer la lisibilité et l'efficacité. Il est également important de constater que nos résultats peuvent constituer un préliminaire à une désambiguïsation automatique du sens de *curieux* en contexte ; comme nous l'avons dit plus haut, les adjectifs situés dans la zone de droite (proches d'*intéressé*) sont les plus éloignés du sens que revêt *curieux* lorsqu'il est antéposé. Le placement d'un substantif dans l'espace de la représentation va permettre de déduire le sens pris par l'adjectif *curieux*. En effet, si le substantif se place au milieu de la zone correspondant à *intéressant*, alors il n'y a pas d'ambiguïté dans le syntagme Nom + *curieux*. C'est le cas des substantifs comme *objet*, *figure*, *type*, *détail*, etc. En revanche, pour les substantifs comme *esprit*, *tête*, *étude*, *bête*, *œil* et *regard*, leur placement s'interprète ainsi : si ces substantifs sont à mi-chemin des deux zones, c'est parce qu'ils possèdent des solidarités lexicales avec les adjectifs de chaque zone, en particulier celle de la zone *intéressé* ; or ces derniers adjectifs étant très peu similaires à *curieux* antéposé, cela veut dire que les sens qu'ils recouvrent ne peuvent être pris que par *curieux* postposé. Par conséquent, c'est l'emploi de *curieux* postposé avec ces six substantifs qui est seul susceptible d'activer ces sens de l'adjectif étudié ; mais cette activation n'est pas systématique, comme le montre l'exemple suivant où il semble que l'auteur ait cultivé à dessein l'ambiguïté du syntagme *regard curieux* :

> « Jacques eut une impression pénible, comme s'il eût essuyé une offense. De minute en minute son ami lui devenait étranger. Un regard curieux, un peu moqueur, dont Daniel l'enveloppa, acheva de le glacer » [R. Martin du Gard, *Les Thibault. Le pénitencier*, p. 792]

Dans cet exemple (analysé dans François, Victorri et Manguin, 2003), l'ambiguïté vient de ce que le co-texte qui précède (*son ami lui devenait étranger*) guide le lecteur vers une interprétation de *regard curieux* en *regard étrange*, tandis que le co-texte qui suit (*un peu moqueur, dont Daniel l'enveloppa*) donne plutôt à *regard* un caractère agentif, et donc une interprétation du syntagme en *regard attentif* ou même *regard inquisiteur*.

Le syntagme *bête curieuse*, habituellement considéré comme figé, illustre le fait que *curieux* ne prend pas toujours le sens *intéressé* quand il est en position postposée ; en effet, dans notre corpus d'étude, nous n'avons pas

trouvé d'exemple où la *bête* en question pouvait être considérée comme agent, et les dictionnaires généraux consultés sur ce sujet confirment ce point de vue. On peut néanmoins citer l'article « curieux » du *Dictionnaire Historique de la Langue Française* : « pour exprimer la valeur passive de « digne d'intérêt, original, étrange » (1559), l'usage a recours à l'antéposition de l'épithète (*un curieux personnage*) afin d'éviter l'ambiguïté, sauf dans *bête curieuse* (1755) » (Rey *et al.*, 1992). La position intermédiaire est donc le reflet d'une ambiguïté possible du syntagme Nom + *curieux*, quand le Nom en question est l'un des six que nous avons mentionnés. Un traitement automatique pourrait ainsi repérer les ambiguïtés possibles et confier la résolution de celles-ci à une exploration du texte avoisinant le syntagme repéré comme éventuellement ambigu. Le placement (autrement dit, le calcul des coordonnées) d'un nouveau substantif, différent des 25 déjà étudiés, se ferait dans cette méthode par le relevé des cooccurrences de ce substantif avec les synonymes de *curieux*, puis par calcul du barycentre comme nous l'avons fait pour les substantifs de l'étude.

8. DICTIONNAIRES SOURCES

Bailly R. 1946. *Dictionnaire des synonymes.* Paris : Larousse.

Benac H. 1956. *Dictionnaire des synonymes*. Paris : Hachette.

Bertaud Du Chazaud H. 1971. *Nouveau dictionnaire des synonymes*. Paris : Robert.

Guizot F. 1864. *Dictionnaire Universel des synonymes de la Langue Française*. Paris : Didier (7ème édition).

Lafaye P.-B. 1858. *Dictionnaire des synonymes de la Langue Française*. Paris : Hachette.

Grand Larousse de la Langue Française. 1971. Paris : Larousse.

Le Grand Robert, dictionnaire de la langue française. 1985. Sous la dir. d'A. Rey. Paris : Robert.

9. RÉFÉRENCES

Berge C. 1958. Théorie des graphes et ses applications. Paris : Dunod.

Borgatti S.P. ; Everett M.G. et Freeman L.C. 1999. UCINET 6.0 version 1.00. Natick MA, Analytic Technologies.

Brodda B. et Karlgren H. 1969. « Synonyms and synonyms of synonyms », SMIL, 5, (p. 3-17). Stockholm.

Coseriu E. 2001. L'homme et son langage. Paris : Peeters.

François J. ; Victorri B. ; Manguin J.-L. 2003. « Polysémie adjectivale et synonymie : l'éventail des sens de curieux », in O. Soutet (ed.) La polysémie. Paris : Presses de l'Université de la Sorbonne.

Frantext, base textuelle catégorisée 1999. CNRS, ATILF (Analyse et traitement informatique de la langue française), UMR CNRS-Université Nancy2, http://www.inalf.fr/atilf

Kahlmann A. 1975. *Traitement automatique d'un dictionnaire de synonymes*. Stockholm : Université de Stockholm.

Kruskal J.B.; Wish M. 1978. *Multidimensional scaling*. Beverly Hills : Sage Publications.

Legendre P. ; Legendre L. 1998. *Numerical Ecology*. Amsterdam : Elsevier.

Manguin J.-L. 2004 (à paraître). « Regroupements de synonymes par indices de similitude : exemple avec l'adjectif *ancien* ». Colloque *Les adjectifs non prédicatifs*, 28-29 novembre 2002, Université de Paris-Nord Villetaneuse. A paraître dans *Les Cahiers de Lexicologie*.

Manguin J.-L. ; Victorri B. 1999. « Représentation géométrique d'un paradigme lexical », *actes de la conférence TALN 1999*, (p. 363-368).

Ploux S. ; Victorri B. 1998. « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *TAL*, Vol 39/1, (p. 161-182).

Rey A. 1992 (éd.). *Dictionnaire historique de la langue française*. Editions Le Robert

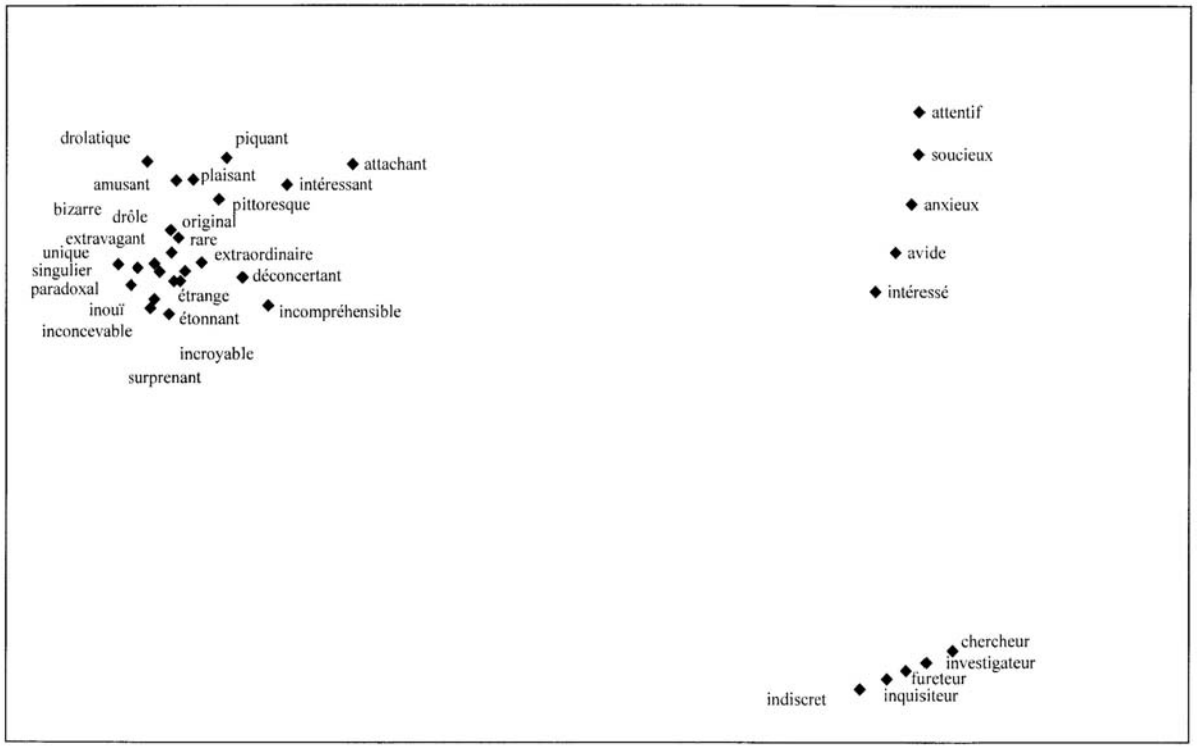


Figure 3 : Représentation de l'espace sémantique de *curieux* par positionnement multidimensionnel, d'après les similitudes paradigmatiques

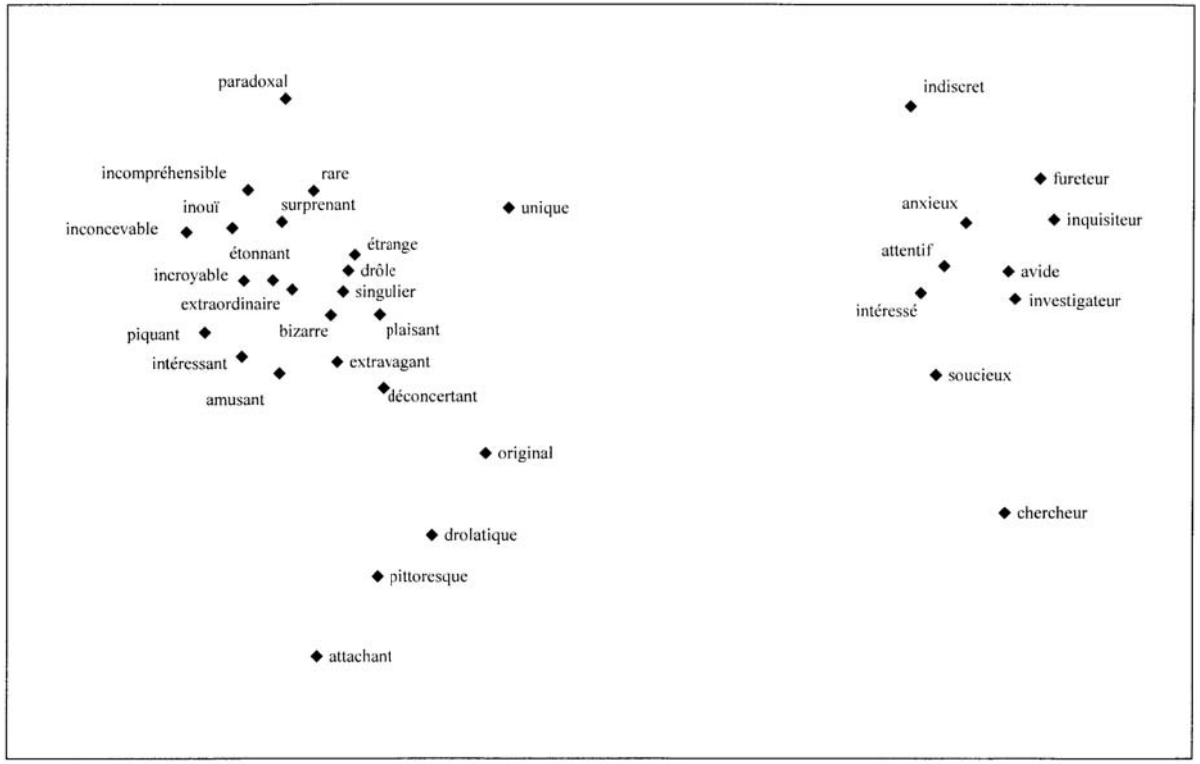


Figure 5 : Représentation de l'espace sémantique de *curieux* par positionnement multidimensionnel, d'après les similitudes d'emploi

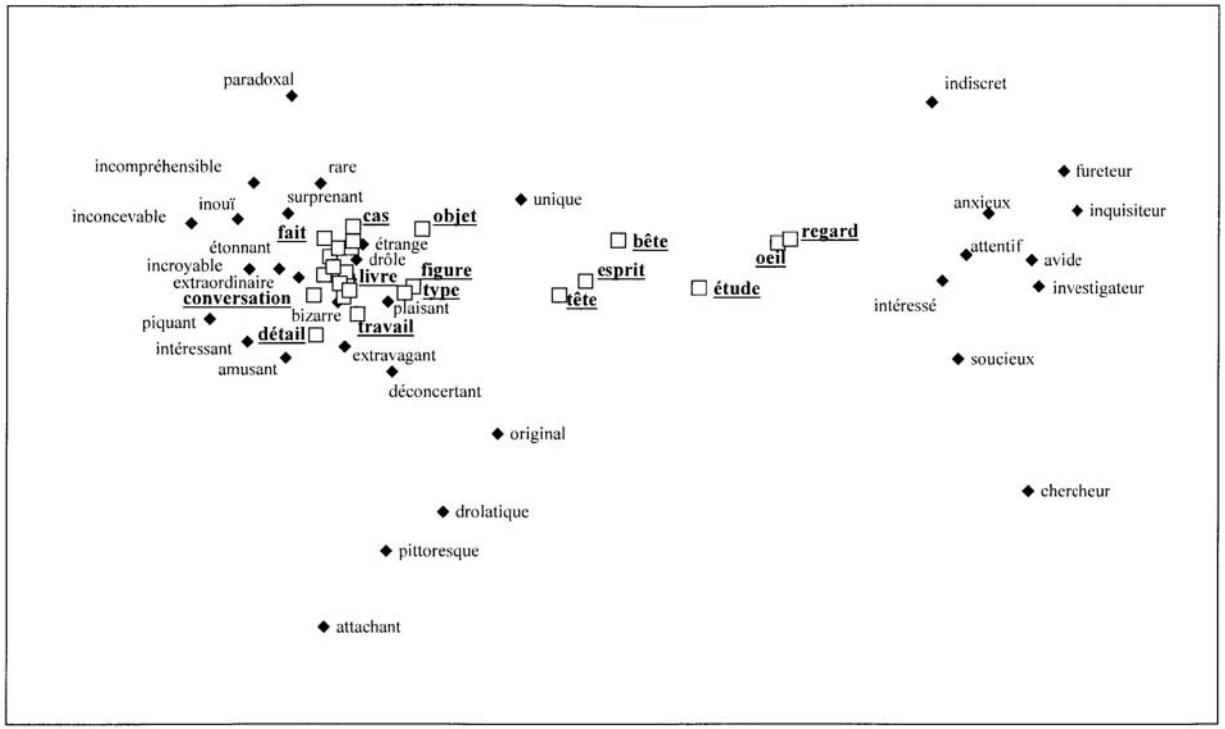


Figure 6 : Représentation de l'espace sémantique de *curieux* et de ses contextes nominaux préférentiels, d'après les similitudes d'emploi

BDVOX : BASE DE DONNÉES POUR SYSTÈMES DE RECONNAISSANCE DE LA PAROLE MULTILOCUTEUR

I. C. Seara, F. S. Pacheco, R. Seara Jr., S. G. Kafka, S. Klein, R. Seara
Université Fédérale de Santa Catarina (Brésil), LINSE

1. INTRODUCTION

La BDVOX est une base de données qui a été créée visant le développement d'un système de reconnaissance de la parole spontanée, indépendant du locuteur et robuste vis-à-vis des variabilités acoustiques et linguistiques, c'est-à-dire le développement d'un système pour la compréhension, à l'aide d'un ordinateur, des mots énoncés par la voix humaine. L'objectif de ces systèmes de reconnaissance de la parole est l'interaction homme/machine, par exemple, dans les services bancaires automatisés, téléphoniques, appels automatiques, services de la voix, agenda personnel, soutien à la liste téléphonique, etc. Les performances d'un tel système dépendent non seulement de la qualité des modèles linguistiques et des paramètres du modèle acoustique, mais aussi de la base de données mise au point pour ce système. Cependant, pour aboutir à un système de reconnaissance de la parole qui soit indépendant du locuteur et de la tâche à remplir, il y a la nécessité d'une énorme quantité de voix, car le système doit s'entraîner le mieux possible. D'autre part, la base de données doit être très bien conçue, pour que tous les sons de la langue à reconnaître, les règles des formations de syllabes et d'autres contraintes linguistiques y soient considérés.

Cette base est une collection de documents sonores qui nécessitent, à partir de ses finalités, une organisation structurée des éléments de la base et des informations associées. La constitution d'une base de données de parole passe par des étapes qui concernent : (i) le choix judicieux du matériau linguistique ; (ii) l'enregistrement sonore ; (iii) la transcription et l'étiquetage. Ce matériau linguistique est constitué d'un ensemble d'échantillons de parole continue et des mots isolés enregistrés par le réseau téléphonique. Chaque extrait possède un ensemble des descripteurs qui contiennent : l'appréciation de la qualité sonore ; la transcription orthographique et phonétique ; des informations relatives aux locuteurs (date

et lieu de naissance) ; des notations des événements prosodiques (point, point et virgule, virgule, deux-points, point d'interrogation, etc.) ; des notations des informations extra-linguistiques (les pauses, les souffles, les rires, les faux départs, etc.). Cette base de données est alors formée de plusieurs *corpus* de la langue, c'est-à-dire d'une collection d'enregistrement de parole qui est accessible à la lecture en ordinateur. Ces données reçoivent des notations et des étiquetages, suffisants pour permettre à un nouveau son un usage dans d'autres phases du système de reconnaissance de la parole.

Étant donné que le but de ce travail est un système multilocuteur, c'est-à-dire les applications « grand public », il faut penser aux différents facteurs qui posent une difficulté à la reconnaissance de la parole. D'abord, il y a la variabilité intra et inter-locuteurs. Après, il faut considérer également les liaisons parce que, dans la parole continue, les mots dans les phrases sont fortement articulés, c'est-à-dire que la prononciation de chaque mot est affectée par le mot qui précède, ainsi que par celui qui suit (Boite *et al.*, 2000). Finalement, il faut observer aussi la variabilité touchant à la lecture ou spontanéité du langage naturel, à savoir si le texte est lu ou parlé librement.

Diverses bases de données ont déjà été établies, toutefois avec des objectifs différents. Elles se distinguent par la langue, le nombre de locuteurs, la forme et les données enregistrées. Voici quelques exemples :

1) EUROM_1 : parole contrôlée par la lecture – 60 locuteurs – salle spéciale – plusieurs langues (cette base de données représente le travail conjoint de huit pays européens : Italie, Angleterre, Allemagne, Hollande, Danemark, Suède, France, Norvège et ultérieurement Grèce, Espagne et Portugal) (Martins *et al.*, 1998) ;
2) BDFALA : parole contrôlée par la lecture – 10 locuteurs – salle à preuve de son – langue portugaise européenne (Martins *et al.*, 1998) ;
3) BD–PUBLICO : parole contrôlée par la lecture – 120 locuteurs – salle à preuve de son – langue portugaise européenne (presque dix millions de mots) (Martins *et al.*, 1998) ;
4) SPEECHDAT : parole contrôlée par la lecture et parlé librement – 5 000 locuteurs – réseau téléphonique – plusieurs langues (Winski, 1997) ;
5) CORAL : dialogues spontanés – 32 locuteurs – salle à preuve de son (Martins *et al.*, 1998)
6) SPEECON : collection de données de parole contrôlée par la lecture et spontanée – au moins 20 langues et 600 locuteurs par langue (adultes et enfants) (Iskra, 2002).

Dans cet article, nous présentons les facteurs qui nous ont conduits au choix du matériau linguistique, soit la méthodologie utilisée pour la création des *corpus* visant l'enregistrement des voix, qui constitueront la base de

données des systèmes de reconnaissance de la parole pour la langue portugaise parlée au Brésil (BDVOX). Ainsi, la BDVOX, développée par l'Université Fédérale de Santa Catarina au Brésil, a été élaborée avec la parole, contrôlée par la lecture et aussi de façon spontanée, enregistrée par le réseau téléphonique. Cette base est composée de 1 100 locuteurs du portugais du Brésil, ce qui fait 193 heures de voix enregistrées. En ce moment, nous sommes en train d'augmenter cette base, en incorporant plus de 500 locuteurs.

2. FACTEURS DÉTERMINANT LA CRÉATION D'UNE BASE DE DONNÉES POUR L'ENTRAINEMENT ET LE TEST DE SYSTÈMES DE RECONNAISSANCE DE LA PAROLE

Comme nous l'avons précisé précédemment, nous allons présenter les procédures utilisées afin d'organiser les divers *corpus* en base de données. Pour commencer ce travail, il faut définir les facteurs déterminant la création de la base de données, quelle que soit la langue à reconnaître. Le système de reconnaissance que nous avons développé est classé pour la reconnaissance, indépendante du locuteur, de mots isolés et enchaînés en phrases courtes ou longues, autrement dit pour la reconnaissance de la parole continue prononcée par un locuteur quelconque. Dans cette optique, il y a la nécessité d'avoir différents *corpus* qui auront divers styles.

2.1. Phrases phonétiquement riches

Le premier facteur considéré est la présence de tous les sons de la langue-but. Dans cette tâche, il faut utiliser de phrases phonétiquement riches, c'est-à-dire un *corpus* qui soit représentatif de tous les sons de cette langue (les phonèmes dans leurs contextes plus fréquents). L'objectif est d'entraîner le mieux possible les phonèmes de la langue en étude, même les plus rares. Il y a là nécessité d'avoir des phonèmes extrêmement rares dans ce *corpus*, parce que son absence dans l'entraînement peut dégrader la reconnaissance des autres phonèmes. Ainsi, chaque locuteur doit fournir de nombreux exemples de chacun des phonèmes du portugais brésilien (dorénavant PB), même les plus rares, comme par exemple, la consonne latérale palatale qui est trop peu fréquente au portugais brésilien (0,21 %) (Seara, 1994). Le Tableau 1 présente l'exemple d'un groupe de phrases phonétiquement riches pour le portugais brésilien qui constitue un des manuscrits gérés par l'ordinateur.

Persistiu o problema na ignição. *Le problème de démarrage a persisté.*	Guto nunca se dedicou a sua etnia. *Guto ne s'est jamais dédié à son ethnie.*
Pneu é um substantivo simples. *Pneu est un substantif simple.*	Usem um manto com florzinhas. *Il faut user un foulard avec des petites fleurs.*
"A Metamorfose" é um livro de Kafka. *"La Metamorphose" est un livre de Kafka.*	Chovendo ou não, irei ao clube. *Pleuvant ou pas, j'irais au club.*
Ele adquiriu um cacto velho. Il a acquis un vieux cactus.	Conte esses bagres. *Comptez ces poissons.*
Untar e pintar são verbos da primeira conjugação. *Graisser et peindre sont des verbes de la première conjugaison.*	O taxímetro registrava o valor errado. *Le taximètre registrait la fausse valeur.*

Tableau 1 : Exemple d'un corpus de phrases phonétiquement riches pour le PB.

Dans ces groupes des phrases phonétiquement riches, il y a les 33 phonèmes du PB (19 consonnes, 2 semi-voyelles, 7 voyelles orales et 5 voyelles nasales) dans les diverses possibilités d'arrangements syllabiques. Pour arriver à cet objectif, il a été construit un logiciel qui fait la vérification de la présence de tous les phonèmes, et calcule les fréquences d'occurrence.

2.2. Parole continue extraite à partir de textes

Le deuxième facteur s'adresse à la parole continue (séquence de mots constituant une phrase). Dans ce cas, il faut chercher des textes ou des paragraphes de différentes longueurs (courts : moins de 20 mots et longs : plus de 20 mots ; extraits des articles de journaux, revues, livres, etc.), pour que le système puisse aussi être entraîné avec les faits prosodiques propres au PB (accentuation, organisation temporelle, rythme et intonation), le changement de rythme des phrases (interrogatives, affirmatives, exclamatives) et le phénomène de la co-articulation, qui consiste en une prononciation, fonction des unités adjacentes (Iskra, 2002). Aussi, dépendant de la quantité de mots qu'il y a dans chaque paragraphe, la lecture est réalisée avec des intonations et une vitesse différentes, paramètres qui affectent le phénomène de la co-articulation. Le Tableau 2 présente des exemples de textes courts et longs, extraits de journaux.

Texte court	**A Telesc está sujeita ao pagamento de multa, em caso de descumprimento ou demora no cumprimento da liminar.** ***La TELESC doit payer d'amendes dans le cas du non accomplissement ou retarde à l'accomplissement du seuil juridique.***
Texte long	A Telesc Brasil Telecom recebeu o prazo de 30 dias para fazer constar, em todas as contas telefônicas dos assinantes e usuários do sistema de telefonia fixa de Santa Catarina, informações detalhadas sobre pulsos e minutos utilizados em ligações locais para telefones fixos, com discriminação de data, horário, duração da ligação e número do telefone chamado, inclusive com relação aos pulsos gratuitos incluídos na franquia. *La Telesc Brasil Telecom doit faire parvenir aux abonnés dans un délai de 30 jours des renseignements détaillées des dates, horaires, durée de l'appel, numéro du téléphone appelé, les impulses et minutes utilisées dans les appels téléphoniques locaux à de téléphones fixes, même pour les impulses qui sont gratuits.*

Tableau 2 : Exemple d'un corpus de parole continue

2.3. Épellation des mots

Un troisième facteur concerne l'épellation des mots. Ce facteur sert à entraîner le système pour des applications nécessitant l'énonciation de noms propres étrangers, comme les noms de famille ou des villes. En effet, la correspondance graphème/phonème dans ces contextes est celle de la langue étrangère concernée. Pour cette raison, ces noms seront épeller. Néanmoins, il y a des personnes qui vont les épeller par syllabes, au lieu de le faire par lettres, ainsi nous attendrons d'avoir dans les enregistrements toutes leurs possibilités d'énonciation. Les mots qui ont été selectionnés pour ce *corpus* composent les groupes des mots orthographiquement équilibrés, c'est-à-dire un *corpus* qui représente tous les graphèmes du PB et dans la même proportion que chacun d'eux apparaît dans la langue (Seara, 1994). Le Tableau 3 présente un *corpus* de mots orthographiquement équibilibrés pour le PB.

Categoria	Catégorie
Problemática	Problématique
Fazendeira	Fermière
Javanês	Langue de Java (île d'Indonésie)
Xadrezista	Joueur d'échecs

Tableau 3 : Exemple d'un corpus de mots orthographiquement équilibrés

2.4. Les nombres

Le quatrième facteur s'adresse aux nombres. Dans ce *corpus*, toutes les façons de dire les nombres (dates, sommes en argent, heures, carte de crédit, téléphone), soient-ils isolés ou en groupes, doivent apparaître. Par exemple, il faut essayer d'avoir toutes les manières de prononcer le numéro d'une carte de crédit. Avec ce critère, nous cherchons à obtenir une grande couverture des possibilités de production de ces items lexicaux. Les consignes employées pour la lecture des nombres sont présentées dans le Tableau 4.

Commande	**Item lu**
Leia este número por extenso *Lisez ce numéro* (*pour les numéros avec 3, 4, 5 et 6 chiffres*)	754.678 (*Ce numéro doit être lu comme : sept cents cinquante quatre mil six cents soixante dix-huit*)
Leia este número *Lisez ce numéro* (*pour les numéros avec un seul chiffre*)	7
Leia esta quantia em dinheiro *Lisez cette somme en argent*	R$ 234,67
Leia este número do cartão de crédito *Lisez ce numéro de la carte de crédit*	1123 2345 4567 6516
Leia esta data *Lisez cette date*	20/06/04
Leia este número de telefone *Lisez ce numéro de téléphone*	2344-3992
Leia a hora *Lisez l'heure*	01:46
Leia esta seqüência de números *Lisez cette sequence de nombres*	51 – 58 – 74 – 26

Tableau 4 : Exemple d'un corpus pour la lecture des nombres

2.5. Parole spontanée

Le cinquième facteur concerne les questions qui amènent à des réponses spontanées, sur le temps ou sur une affaire quelconque. Le terme « spontanées » sert à désigner les productions orales dans lesquelles le sujet élabore son message à l'instant où il le produit. Il y a de grandes différences entre les textes lus et ceux parlés spontanément. Ces différences sont dans l'intonation, et même dans les pauses qui sont faites et les répétitions, que les gens font quand ils ne sont pas en train de lire, c'est-à-dire des phénomènes propres à la parole spontanée (hésitations, respirations, reprises, syntaxe de l'oral, etc.).

D'ailleurs, comme la plupart de ces systèmes de reconnaissance de la parole servent à l'exécution de commandes à travers les réseaux téléphoniques, il a été important aussi d'enregistrer un *corpus* avec des mots de commande, par exemple, *décrocher*, *appel*, *temps*. Pour avoir toutes les possibilités de variation d'intonation, ces mots de commande ont été introduits dans des phrases qui font les demandes, par exemple : *s'il vous plaît*, *vous pouvez faire l'**appel ?*** ou *s'il vous plaît **décrocher** l'appareil* ou *quel sera le **temps** demain ?* Ces mots de commande doivent être choisis pour envisager le plus grand nombre d'application.

De plus, dans plusieurs applications des systèmes de reconnaissance de la parole, les mots à reconnaître sont OUI et NON. Ils servent à confirmer, ou pas, les résultats des commandes vocales. De cette façon, il a été conçu un *corpus* de questions ayant comme réponse OUI/NON. Dans ce cas, nous essayons de trouver un équilibre entre les questions qui auraient les réponses OUI et celles avec les réponses NON (Tableau 5).

Questions	**Réponses possibles**
Você está falando de um telefone fixo? *Vous êtes en train de parler par un téléphone fixe?*	OUI/NON
Você está falando de um telefone celular? *Vous êtes en train de parler par un téléphone mobile?*	OUI/NON
Você é do sexo feminino? *Vous êtes du sexe féminin?*	OUI/NON
Você é do sexo masculino? *Vous êtes du sexe masculin?*	OUI/NON

Tableau 5 : Exemple d'un corpus pour avoir des réponses OUI/NON

3. LES ENREGISTREMENTS

En ce qui concerne les applications des systèmes de reconnaissance de la parole multilocuteurs, il nous semble que les différents *corpus* présentés dans cette étude sont indispensables. De plus, les enregistrements ont été faits dans les environnements qui avaient une très grande variabilité acoustique (bruit de fond, présence de musique, parole superposée, etc.), circonstance idéal pour ces enregistrements, car le système doit être capable de fonctionner proprement dans des conditions difficiles.

Chaque *corpus* envisagé a été construit avec un grand nombre de groupes représentatifs. Par exemple, le *corpus* des phrases phonétiquement riches est constitué de 80 groupes comportant dix phrases chacun. Avec les *corpus* (mots phonétiquement riches, mots à épeler, les nombres, les questions OUI/NON, les réponses spontanées), nous avons créé des manuscrits. Ces manuscrits sont composés par un groupe représentatif de chaque *corpus* qui est sélectionné de façon ordonnée. Tous ces manuscrits ont été enregistrés par le réseau téléphonique fixe. À l'heure actuelle, nous avons déjà enregistré 1 100 locuteurs.

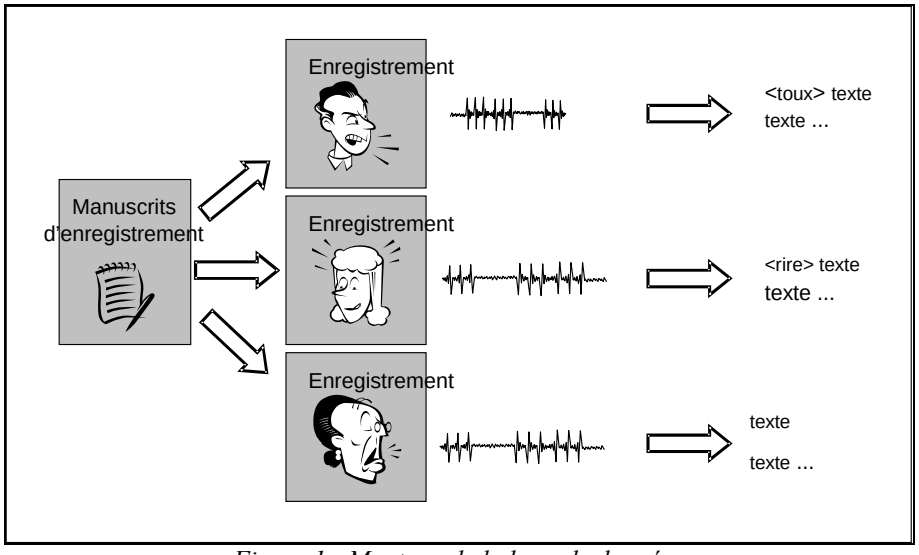


Figure 1 : Montage de la base de données

Ainsi, la BDVOX est une base de données constituée de 27 572 mots différents, qui comptent 1 491 512 mots si l'on considère tous les mots des manuscrits créés à partir des *corpus*. Cette base a été transcrite orthographique et phonétiquement. Nous avons fait aussi la transcription de tous les événements acoustiques, comme les bruits de respiration, les toux, les bruits de lèvres, le rire, les pauses longues mises entre les mots et aussi

les bruits de l'environnement, quand ils sont vraiment remarquables. Chaque enregistrement a eu une durée de presque 10 minutes.

La BDVOX fournira les bases pour les tests et les entraînements de systèmes de reconnaissance de mots isolés, de la parole continue, de commandes pour les téléservices. Les Tableaux 6, 7 et 8 présentent la composition actuelle de la BDVOX par rapport aux statistiques générales, la distribution des phrases par rapport au sexe et l'âge, et la qualité d'enregistrement.

Statistiques du BDVOX	
Nombre de phrases	216 560
Nombre d'heures	193,73
Nombre de mots	1 491 512
Nombre de mots différents	27 572

Tableau 6 : Statistiques générales

Sexe	Nombre de phrases	Nombre d'heures	Pourcentage dans la base
Masculin	134 538	120,70	62,30%
Féminin	81 909	72,98	37,67%
Âge			
Enfant	3 303	2,62	1,35%
Adulte	211 427	189,61	97,87%
Âgé	1 717	1,44	0,75%

Tableau 7 : Distribution des phrases par rapport au sexe et l'âge

Bruit de Fond	Nombre de phrases	Nombre d'heures	Pourcentage dans la base
Perceptible	39 750	39,81	20,55%
Pas Perceptible	176 696	153,87	79,42%
Qualité perçue			
Mauvaise	9 753	8,52	4,40%
Moyenne	24 461	25,69	13,26%
Bonne	170 968	149,04	76,93%
Excellente	9 262	10,42	5,38%

Tableau 8 : Qualité des enregistrements

4. CONCLUSION

La BDVOX s'est montrée être une base de données très robuste, et aussi un outil très riche et fondamental au développement des systèmes de reconnaissance de la parole pour le portugais parlé au Brésil. Maintenant, nous sommes en train d'élaborer une deuxième étape, pour augmenter le nombre des locuteurs de la base de données. Ainsi, nous ajoutons plus de 500 nouveaux locuteurs, en essayant d'équilibrer le pourcentage des locuteurs du sexe masculin et du sexe féminin. Dans la première étape, il y a eu un fort biais sur le nombre des locuteurs masculins (voir Tableau 7).

5. RÉFÉRENCES

Martins, C., Mascarenhas, C. I., Meinedo, H., Neto, J. P., Oliveira, L., Ribeiro, C., Trancoso, I., Viana, M. C. 1998. "Spoken Language Corpora for Speech Recognition and Synthesis in European Portuguese". *Proc. of the 10th Conference on Pattern Recognition*, RECPAD'98, Lisbon, mars 1998, p. 357-364.

Winski, R. 1997. "SPPECHDAT : Definition of corpus, scripts and standards for Fixed Networks" *Tech. Rep. LE2-4001-SD1.1.1*., jan. 1997.

Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., and Kiessling, A. 2002. "SPEECON - Speech databases for consumer devices : Database specification and validation" *Proc. 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, mai 2002, p. 569-572.

Boite, R., Bourlard, H., Dutoit, T., Hancq, J., Leich, H. 2000. *Traitement de la parole.* Lausanne : Presses Polytechniques et Universitaires Romandes.

Combescure, P. 1981. "20 listes de dix phrases phonétiquement équilibrées". *Revue d'Acoustique*, n° 56, jan 1981, p. 34-38.

Seara, I. C. 1994. *Etude Statistique des phonèmes du portugais parlé dans la capitale de l'État de Santa Catarina pour l'élaboration de phrases phonétiquement équilibrées (en portugais)*, Master of Science, UFSC, Florianópolis, Brasil

.

CORPUS, VOUS AVEZ DIT CORPUS ! DE LA NOTION DE CORPUS À LA CRÉATION D'UN « CORPUS INFORMATISÉ »

Céline Vaguer
UMR 7114 – MoDyCo – Université Paris X-Nanterre

1. INTRODUCTION

On ne peut mener un travail linguistique sans référence à des « données » : ainsi toute grammaire ou tout dictionnaire arbore des « exemples » ; on ne parle pas pour autant, dans ces cas, de « corpus » : il semble que la notion soit entendue (en particulier dans les débuts de la grammaire générative) comme « un ensemble de données produit indépendamment du linguiste et de la recherche linguistique », par opposition aux données que le linguiste est susceptible de produire lui-même : Chomsky s'oppose à l'idée que l'étude d'un corpus puisse mener à la construction d'une grammaire appropriée, comme à l'idée que le corpus des énoncés que l'enfant entend autour de lui soit la base de sa « compétence » (de la grammaire qu'il se construit mentalement). Ainsi le raisonnement linguistique de Chomsky s'opère bien sur des données concrètes, mais qu'il n'appelle pas « corpus ». Donc le débat instauré par Chomsky – étant donné le sens qu'il donne à « corpus » comme, disons, un « ensemble de discours produit extérieurement au linguiste et au travail linguistique » – c'est celui de la pertinence du « corpus » par rapport à ce que le linguiste (de par sa compétence de sujet parlant) peut produire lui-même, d'une part, ou par rapport à ce que la grammaire telle qu'il l'a construite peut prédire, d'autre part. L'argument de Chomsky à l'encontre du « corpus » (comme base pertinente de la description et du raisonnement linguistique), c'est le fait que, pour raisonner sur la langue, il faut pouvoir confronter ce qui est possible et ce qui ne l'est pas, or – par définition – le « corpus » (tel qu'il l'entend) ne peut pas fournir d'exemples de ce que la langue ne permet pas ; de plus, en tant que texte produit à un moment donné, par un ou des locuteurs particuliers, selon un thème, une intention, une situation, des interlocuteurs particuliers, un « corpus » ne peut évidemment illustrer tous les cas de figure d'un phénomène linguistique donné (par exemple : tous les auxiliaires et

combinaisons d'auxiliaires) ; et enfin, en tant que produit fini, le « corpus » ne peut pas non plus laisser voir certaines propriétés linguistiques comme la récursivité (le retour potentiellement infini d'une même structure).

Le présupposé est que le linguiste, de par sa propre compétence de sujet parlant, est à même de produire les données pertinentes (grammaticales et agrammaticales), permettant de faire l'hypothèse de règles dont il vérifiera la pertinence en jugeant si l'ensemble des énoncés qu'elles peuvent produire est, ou non, conforme à ce qu'autorise la langue – c'est-à-dire ce que le linguiste lui-même considère comme acceptable ou inacceptable. Ainsi, dans ce cadre, le travail du linguiste suppose nécessairement le recours à l'intuition pour constituer les données, les manipuler, raisonner sur le résultat de ces manipulations, mais en même temps, il y a un doute sur la pertinence de l'exercice de l'introspection – ce pourquoi justement les structuralistes et les distributionnalistes avaient prôné le recours au « corpus ». Mais on sait aussi que ce dernier n'est pas la panacée, ainsi que l'a pointé Chomsky. Toute recherche entreprise doit donc se mettre au clair sur ce point méthodologique :

- qu'est-ce qu'un corpus ?
- quel est ou quel doit être le statut du corpus dans l'investigation linguistique ?

Dans un premier temps, nous ferons un bilan sur cette notion de *corpus* – bilan né du constat que bien souvent, dans les articles de linguistique, rien n'est dit par les linguistes sur le statut des données : « ressources dont les natures différentes ne sont pas nécessairement distinguées par le linguiste, qui les nommera toutes "corpus" » (Gasiglia, 2003), mais aussi du fait que les supports de recherche d'occurrences ont évolué et qu'il est donc primordial de réfléchir sur la nature des données ainsi récoltées. Pour ce faire, nous mettrons en évidence l'existence de différentes conceptions de la notion de *corpus*, de différentes attitudes à l'égard des données, de différentes démarches pour élaborer les corpus, de différents jugements que l'on produit sur les données. Puis, nous justifierons le point de vue que nous avons adopté en tant que chercheur, et nous exposerons la démarche retenue pour constituer notre corpus : la méthodologie et la constitution d'une base de données.

2. LA NOTION DE *CORPUS*

Quelles que soient la théorie et la méthodologie retenues, se pose à tout linguiste la question de la définition du corpus puisque c'est ce dernier qui l'amène à pouvoir formuler une hypothèse ou à en éprouver la consistance. Saussure (1916 *in* 1972) avait raison de dire que « en matière de langue, on s'est toujours contenté d'opérer sur des unités mal définies ».

2.1 Les différentes conceptions de la notion de corpus

L'existence de différentes conceptions de la notion de *corpus* apparaît lorsque l'on regarde comment les linguistes l'abordent et la définissent. Pour les uns, il faut entendre par là un ensemble d'énoncés retenus, écrits ou oraux (parmi l'*univers*[1] des possibles), qui sera soumis à l'analyse : « base d'observation permettant d'entreprendre la description et l'analyse de la langue en question » (Arrivé *et al.*, 1986). Mais pour d'autres, le corpus est en fait issu d'un travail préalable, puisque l'ensemble est restreint à ce qui est considéré comme « représentatifé » ; c'est le cas de Riegel *et al.* (1994) qui spécifient de surcroît que les données doivent être « attestées » :

> « On peut rassembler un ensemble de textes ou d'énoncés jugés représentatifs de la langue… Une telle collection ne comprenant que des données attestées (des énoncés effectivement produits) constitue un *corpus* ».

Le corpus retenu, qui aura alors subi un jugement d'acceptabilité de la part du linguiste, puisque « le linguiste trie les énoncés qu'il va soumettre à l'analyse » (Dubois *et al.*, 1999), sera considéré comme un « échantillon de la langue » (*op. cit.*) que tout linguiste souhaite représentatif[2], en ce sens qu'il espère qu'il illustre l'ensemble des possibilités structurelles existantes (par exemple de l'emploi de la préposition *dans*), tout en sachant qu'il ne sera pas exhaustif[3] puisqu'on ne peut prétendre rassembler tous les énoncés possibles…

2.2 Les différentes attitudes à l'égard des données

Ainsi, existe-t-il autant de corpus que d'objets d'étude, mais aussi autant de corpus que de points de vue non seulement théoriques et méthodologiques, ou encore selon que l'on est lecteur ou chercheur (Vaguer, 2004b & 2005b). On peut, en effet, retenir le point de vue du lecteur, qui prend connaissance d'un certain travail, d'une part, et le point de vue du chercheur qui opère le travail en question ; les deux « corpus » ainsi délimités ne se recoupent que partiellement : si nous nous définissons en tant que lecteur, le corpus de Vandeloise (1986), par exemple, correspond alors à l'ensemble des phrases constituant l'objet de l'analyse présenté dans l'ouvrage, mais ce n'est sans doute qu'un sous-ensemble (celui que l'auteur a retenu comme pertinent pour l'exposé) de la totalité des exemples effectivement examinés par Vandeloise ; c'est ainsi que Milner (1978) peut écrire :

> « Les exemples, comme il est d'usage dans la grammaire transformationnelle, sont censés valoir pour la classe entière des phrases construites de manière analogue. De façon générale, nous laisserons à l'intuition du lecteur le soin de reconstituer la classe pertinente. »

2.3 Les différentes démarches pour élaborer les corpus

Si l'on adopte le point de vue du chercheur, il y a à nouveau à distinguer entre deux démarches possibles (Fillmore, 1992)[4] : ou bien les hypothèses s'élaborent à partir d'exemples « forgés » (l'« introspection » dans le cadre d'une « linguistique de bureau », Corbin 1980), ou bien le travail s'opère sur des exemples « attestés » (le « corpus » dans le cadre d'une « linguistique de terrain », *Ibidem*) ; dans le premier cas, le linguiste construit lui-même les énoncés, dans le second cas, il les relève dans des textes de divers genres qui n'ont pas été produits pour les besoins de la cause (romans, articles de presse, entretiens radiophoniques, etc.).

2.3.1. CORPUS FORGÉ : AVANTAGES ET DÉSAVANTAGES. L'une des façons pour un linguiste de constituer les données, sur lesquelles il va travailler, repose sur ce que l'on appelle « les corpus forgés » : « corpus basés sur la pratique expérimentale et dynamique qui consiste à utiliser la compétence des locuteurs pour obtenir des données selon les besoins de l'étude » (Riegel *et al.*, 1994). Le linguiste peut alors s'adresser à des informateurs[5] pour savoir quels sont leurs jugements d'acceptabilité sur l'ensemble des énoncés, pour leur faire produire des énoncés et ainsi vérifier la représentativité de ses propres réactions.

a) Le principal avantage de l'exemple forgé est qu'il permet les manipulations dont le linguiste a besoin pour procéder à son analyse et observer celles qui ne sont pas possibles[6] (éventualité peu probable dans les énoncés attestés). Soit, par exemple, l'énoncé *Il est dans les dix heure*s : il peut être soumis à diverses commutations permettant de conclure, rapidement et économiquement (par rapport au temps que représenterait la recherche effective des phrases attestées correspondantes), que la préposition peut se voir substituer *vers* mais non *à, de, pour*…, et que le déterminant est incommutable [**Il est dans (ces + mes + des + quelques + plusieurs) dix heures*]. De même, si l'on cherche quels compléments de verbe *dans* peut introduire, plutôt que de procéder à des relevés dans des textes, il est peut-être plus sûr de tester à partir de la liste fournie par un dictionnaire quels verbes sont susceptibles de se construire avec *dans*, et quelles sont les propriétés permettant de les classer… L'intérêt de cette démarche est qu'elle est relativement objective parce qu'indépendante des aléas des corpus attestés (on peut avoir en effet un article de presse ou une page de roman sans un seul complément en *dans* – *a fortiori* un complément de type précis que l'on cherche à étudier). De plus, les corpus forgés ne nécessitent pas de longues et fastidieuses manipulations d'exemples, tel que c'est le cas avec les corpus attestés où les phrases sont généralement plus longues et complexes. La constitution de corpus forgés s'avère alors plus souple et plus

économique (en temps et en investissement notamment) que le dépouillement de corpus divers.

b) Le principal désavantage de l'exemple forgé est qu'il est tributaire des jugements d'acceptabilité et de grammaticalité du chercheur (nous reviendrons plus loin sur ces notions), et que ces derniers peuvent être faussés (involontairement) par la prégnance de l'hypothèse que l'on a en tête ; ainsi Melis (2003) considère t-il que *dans les* ne peut introduire un sujet (il met l'astérisque à **Dans les deux cents kilos suffiront* et **Restent dans les trente semaines à planifier*), alors que le lecteur forgera facilement (d'ailleurs précédé par Gross, 1977) entre autres *Dans les trente personnes sont venues*, ou acceptera les énoncés incriminés. De plus, étant donné que nul n'est parfait, on n'est jamais sûr de penser à toutes les possibilités qu'offrent tous les items, et on peut fausser les tests (plus ou moins consciemment) en fonction de l'hypothèse qui se fait jour. Enfin, un autre désavantage des corpus forgés est qu'ils ne permettent pas de décrire (qualitativement et quantitativement) la représentativité des données dans l'usage effectif de la langue : « l'introspection est impuissante à décrire leur [les variations dans les pratiques langagières] distribution dans la population : le social lui échappe par définition » (Corbin, 1980).

2.3.2. CORPUS ATTESTÉ: AVANTAGES ET DÉSAVANTAGES. Les corpus attestés se définissent par le fait que les données ont été produites indépendamment du travail linguistique, qu'elles relèvent de sources diverses (romans, article de presse, etc.) et qu'elles peuvent être de natures diverses (écrites ou orales).

a) Les corpus attestés présentent certains avantages (par rapport aux phrases forgées) : l'auteur (du roman, de l'article de presse, etc.) fait un usage spontané de tel terme ou de telle structure ; il n'y a donc pas de risque que la phrase qu'il produit soit faussée par une hypothèse (d'ordre linguistique) à démontrer : les données n'ont pas été produites pour les besoins de la recherche linguistique, ni suscitées par elle. Elles n'ont ainsi pas subi l'influence du linguiste (comme cela peut se produire lorsqu'il forge ses exemples).

b) L'utilisation de corpus attestés présente toutefois des désavantages : un corpus – si vaste soit-il – ne comporte pas nécessairement toutes les données pertinentes (par exemple toutes les manipulations permettant, dans la suite Verbe + Infinitif, de distinguer entre semi-auxiliaire (*Il va partir*) et verbe distributionnel (*Il désire partir*)). En revanche, on peut y trouver des cas de figure auxquels on n'aurait pas pensé spontanément. Il faut dire aussi qu'un cas de figure représenté dans un corpus attesté peut tout simplement ne pas être remarqué par le chercheur : il y a une longue tradition grammaticale et lexicographique qui s'appuie sur des exemples attestés mais qui, entre

autres, n'a jamais repéré certains emplois de *dans* ; ainsi tous les dictionnaires signalent-ils le sens spatial, le sens temporel, le fait que *dans* puisse introduire un état (*être dans l'embarras*) ou l'approximation (*Il a dans les trente ans*), mais aucun ne mentionne l'interprétation appositive (Leeman 2000 ; Vaguer 2000) que peut prendre *dans ce tableau, JE vois dans ce tableau une preuve de sa folie*, compris comme « Ce tableau est une preuve de sa folie ». Le recours à des corpus attestés ne garantit donc pas à lui seul la complétude ou la représentativité de la description. De plus, l'objectivité qu'ils procurent n'est pas entière. Si le corpus c'est, par exemple, la liste des compléments en *dans* que l'on peut extraire de *Frantext*, c'est un recensement neutre. Ce qui n'est pas neutre, c'est ce que l'on fait de ce recensement : on va opérer une sélection selon ce que l'on cherche à étudier, par exemple les compléments temporels ; on s'éloigne de l'objectivité dans la mesure où c'est le linguiste qui décide de ce qui est (ou non) temporel, et donc fait intervenir une certaine intuition (par conséquent nécessairement une certaine subjectivité) – même s'il applique des critères, le résultat qu'il affecte au test dépend de son sentiment linguistique. Dans ces compléments temporels, on ne va en garder qu'un certain nombre, sur la base là aussi de jugements personnels : on élimine ce qui paraît redondant, du même type ; on garde ce qui semble le plus propre à illustrer ce que l'on veut dire, mais on ne signale pas ce sur quoi on n'a rien de particulier à observer, etc.

2.3.3. CONCLUSION. Dans les deux cas, donc, il y a le risque que le chercheur manque des données pertinentes, du fait que, aussi bien lorsqu'il forge des phrases que lorsqu'il recherche des énoncés attestés, il est plus ou moins inconsciemment guidé par une certaine chose à découvrir, ce que masquent les formulations passives dans les définitions habituellement fournies du corpus. Ainsi, pour Arrivé *et al.* (*op. cit.*) c'est un « ensemble d'énoncés d'une langue donnée (écrits ou oraux enregistrés) qui ont été recueillis pour constituer une base d'observation permettant d'entreprendre la description et l'analyse de la langue en question » – où rien n'est dit sur les critères qui président au « recueil »[7]. Et si des critères sont précisés par Sinclair (1996) : « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage », ils le sont sur la base de principes avouables (et même garants de scientificité), effaçant tous les risques liés à la subjectivité du linguiste ! Pour Sampson (1994), « la linguistique de corpus prend le langage comme elle le trouve ». Or, on ne prend pas le langage tel qu'il est, même dans une linguistique de corpus, à partir du moment où l'on interprète nécessairement les énoncés (attestés) – ce dont témoignent précisément les différences d'acceptabilité[8]. Ainsi, quand on relève des énoncés, on les trouve attestés par rapport à un contexte donné. Or

qu'en est-il quand les données sont appréhendées hors contexte, par d'autres lecteurs ? Il se peut que ces derniers, confrontés à des phrases isolées, puissent être dans l'incapacité de trancher sur leur acceptabilité ou puissent leurs attribuer un jugement d'acceptabilité différent. La notion d'acceptabilité est fluctuante dans la langue ; on n'est donc jamais sûr que ce qu'une autre personne qualifie d'acceptable le soit pour les mêmes raisons, selon les mêmes facteurs que soi. C'est en cela que l'établissement d'un corpus est toujours marqué de subjectivité car, qu'on le forge ou qu'on le relève, c'est toujours suivant ses propres intérêts de recherche, en ayant déjà une idée dans certains cas du type de structure que l'on cherche. L'objectivité revendiquée par les tenants du corpus attesté n'est qu'apparente, cachant un jugement d'acceptabilité refoulé.

2.4 Les différents jugements que l'on produit sur les données : l'acceptabilité et la grammaticalité

Le jugement que porte le linguiste sur les énoncés est le seul outil dont il dispose pour séparer, dans les données qu'il observe ou qu'il construit, celles qui peuvent fonctionner comme illustration de ce que la langue permet, de celles qui montrent ce que la langue interdit ; c'est à partir de cette base fondamentale que l'on peut saisir les différences entre les deux ordres d'énoncés qui vont justifier la formulation d'une règle : le fait que l'on puisse affirmer qu'en français le déterminant se trouve devant le nom repose sur l'observation que, si est possible (entre autres) *Le soleil brille*, ne le seraient pas *Soleil brille* ni *Soleil le brille* ni *Soleil brille le*. Cependant, pour étiqueter ces possibilités et impossibilités, deux termes existent : les uns parlent de « grammaticalité », les autres d'« acceptabilité », d'autres encore des deux.

2.4.1. Ainsi pour Milner (1978), « le jugement d'acceptabilité est le seul donné sur lequel le linguiste peut raisonner pour construire, en terme de grammaticalité, sa théorie ». Mais ce jugement n'est pas objectif, en ceci qu'il n'est pas porté pour constater le résultat d'une expérimentation mettant en jeu des outils indépendants de lui (comme dans le cas du chimiste qui constate que l'addition de tel acide dans telle solution la fait virer au bleu, ou a pour effet un bouillonnement, etc.) – d'où les tentatives de saisir ce qui est en jeu dans le jugement afin de l'objectiver – d'en faire la base d'un critère reproductible – du fait que les locuteurs n'ont pas forcément les mêmes réactions face à un même énoncé.

2.4.2. Pour Picabia & Zribi-Hertz (1981), « sera dite *grammaticale* dans la langue L, une séquence conforme aux principes et contraintes linguistiques qui constituent la *grammaire* de L ». La grammaticalité relève donc de la compétence. La définition de Picabia & Zribi-Hertz semble totalement

circulaire puisque pour constituer une grammaire, on se fonderait sur le jugement de grammaticalité, et que la grammaticalité, c'est le jugement que les phrases appartiennent à la grammaire ! Autrement dit, on retombe dans le problème posé par le recours à l'acceptabilité : pour élaborer une grammaire, on a besoin de savoir ce qu'est une phrase grammaticale, avant même que la grammaire soit élaborée ; sur quelle base alors décide t-on que la phrase est (a)grammaticale ? Sur une intuition – dont les conditions d'exercice ne sont pas claires : la preuve, on ne fait pas de partage net entre acceptabilité et grammaticalité. Cette confusion (relevée par Normand, 1972) explique que l'on ait souvent reproché aux générativistes de se fonder en fait sur un sentiment linguistique reposant sur la norme (traditionnelle : le « bon » usage) de l'écrit : il est évident que l'on ne construira pas les mêmes règles disant ce qui appartient à la grammaire (qui se confond en l'occurrence avec la langue elle-même), selon que l'on part, par exemple dans le cas de l'interrogation, d'un corpus rassemblant comme phrases grammaticales (à l'exclusion des autres, jugées agrammaticales) :

- ou bien : *Où vas-tu ?* et *Où est-ce que tu vas ?*
- ou bien : *Où vas-tu ?*, *Où tu vas ?*, *Tu vas où ?* et *Où est-ce que tu vas ?*, *Où c'est que tu vas ?*, *C'est où que tu vas ?*

(dans le premier cas, mais non dans le second, *Où tu vas ?* et *C'est où que tu vas ?* entre autres seront jugées non conformes à la grammaire).

2.4.3. Mais pour beaucoup de linguistes, définir cette notion de grammaticalité ne peut se faire sans passer par celle d'acceptabilité, elle en serait d'ailleurs une partie (dans l'acceptable, il y a du grammatical) : ainsi pour Riegel *et al.* (1994) « la grammaticalité ne regrouperait que la partie de l'acceptabilité qui est déterminée par les règles de bonne formation intrinsèque des énoncés », ou chez Picabia & Zribi-Hertz (1981) « la grammaticalité est la composante linguistique de l'acceptabilité ». En fait, l'utilisation des termes paraît renvoyer à des niveaux différents : dans le cas de Milner, le jugement intuitif est dit d'« acceptabilité » ; c'est celui qui permet de trier les formes et d'élaborer par hypothèse une grammaire, laquelle produira des formes (dites, donc, « grammaticales ») ; dans le cas de Riegel *et coll.*, la grammaticalité relève de la structure, tandis que l'acceptabilité a trait aux compatibilités distributionnelles : *Le soleil nage* est grammatical mais inacceptable, *Soleil le brille* est agrammatical. Le problème est dans la circularité de la démarche : si l'on juge *Est-ce que le soleil brille-t-il ?* inacceptable, on construira une grammaire de telle sorte qu'elle ne produise pas cette séquence (dite, donc, agrammaticale).

3. LA CRÉATION D'UN « CORPUS INFORMATISÉ » : UNE BASE DE DONNÉES LINGUISTIQUE

3.1 Corpus forgé ou corpus attesté ?

Dans le cadre de notre recherche, centrée sur la préposition *dans*, corpus forgé et corpus attesté sont complémentaires et non concurrents. Les énoncés attestés viendront de sources diverses (essentiellement écrites) et les énoncés forgés émaneront des manipulations produites sur ces données attestées.

3.1.1. Les phrases forgées permettent le test rapide et économique des propriétés (que l'on souhaite aussi représentatif que possible), par exemple concernant l'association d'un verbe à un complément en *dans*, et des propriétés syntaxiques que possèdent l'énoncé ainsi construit (par exemple *dans la fuite* n'est ni supprimable ni déplaçable dans *La solution est dans la fuite*). De plus, elles permettent de pallier les « trous » éventuels (ou inéluctables) des corpus attestés (ainsi, il y a peu de chance *a priori* que l'on puisse constituer, à partir des corpus attestés disponibles, la liste des verbes susceptibles de se construire avec *dans*), et de construire des associations agrammaticales ou inacceptables qui, comparées aux suites recevables, sont susceptibles de donner des idées d'hypothèse pour caractériser le complément dont on s'occupe. Notre objectif premier est d'étudier la langue, c'est en cela que nous nous rapprochons davantage du champ harrissien et que nous nous éloignons du champ chomskyen.

3.1.2. Les extraits attestés permettent de vérifier ou d'amender les propositions de description ou d'explication, avancées à partir de corpus forgés, et d'en pallier les manques (en portant à l'observation des énoncés auxquels le chercheur ne pense pas spontanément), donc de limiter le risque de circularité (lorsque le chercheur muni d'une certaine hypothèse secrète les observables qui vont dans le même sens). En effet, ainsi que le signale Blanche-Benveniste (2000), « comme ils [les corpus] contiennent des données attestées, dont on peut vérifier les sources, ils engagent à faire un travail d'analyse linguistique qui ne repose pas uniquement sur l'intuition mais sur la confrontation avec des données parfois étonnantes, que la simple intuition n'aurait pas pu atteindre ».

3.1.3. L'objectif de notre recherche n'est pas de quantifier les emplois de la préposition *dans*[9], mais bien d'avancer dans son identification syntaxique et sémantique. Notre objectif est donc descriptif et concerne la caractérisation de *dans* en langue, et non pas en discours (c'est-à-dire dans les productions orales ou écrites telles que rassemblées dans les corpus attestés) ; autrement dit, il ne s'agit pas de voir comment les locuteurs utilisent *dans* (ou tel type

de complément en *dans*) : plutôt à l'oral qu'à l'écrit ou inversement, plutôt dans la description que dans la narration ou l'argumentation, ou réciproquement, plutôt dans tel genre que dans tel autre, etc., ni donc de voir quel emploi est le plus représentatif ou le plus fréquent dans les performances. Il s'agit de déterminer à quels différents emplois de *dans* on a affaire dans les discours (seuls observables : les actualisations de la langue sont le passage obligé de tout travail linguistique, comme on l'a vu précédemment), de façon à essayer de construire une identité de la préposition en langue permettant, en retour, de rendre compte des énoncés concrets dans lesquels elle apparaît. Le corpus est donc une base incontournable : ce à partir de quoi on peut avoir un aperçu des différentes possibilités qui guident la recherche d'une définition, ou qui en permettent la vérification lorsqu'on a élaboré une hypothèse, mais qui n'est pas en lui-même l'objet de la recherche (notre objectif n'est pas l'analyse des discours). C'est en cela qu'on se rapproche de la « linguistique de corpus » entendue comme : (a) « le travail que fait le linguiste qui constitue un corpus », c'est-à-dire qui prend un texte (au sens large : écrit/oral transcrit, etc.), l'annote par l'ajout d'informations d'ordre morphologique, syntaxique, sémantique… et le traite informatiquement (étiquetages, arbres, analyseurs syntaxiques) pour le rendre utilisable par d'autres (outil d'exploration) puisque nous constituons un corpus (informatisé), et non en tant que (b) le corpus serait l'objet même de notre étude – puisque ce qui nous intéresse c'est un « fait de langue ». En effet, si l'on se reporte à l'opposition saussurienne langue/parole, reformulée en langue/discours, le corpus tel que défini en (a) est un discours (traité informatiquement), qui nous intéresse en tant qu'il manifeste des emplois (effectifs), en tant qu'il témoigne d'emplois possibles (attestés). Mais notre objectif n'est pas de rendre compte des emplois dans tel ou tel corpus (si étendu soit-il) : on cherche à saisir l'identité de la préposition *dans* en langue, identité formelle et sémantique censée présider aux / déterminer les multiples actualisations en discours. Donc le corpus n'est pas l'objet même de notre recherche (puisqu'on ne cherche pas à décrire un corpus), il n'en est que l'outil (incontournable, certes).

3.2 La démarche adoptée

En ce qui concerne la complémentation verbale, la démarche de constitution du corpus a consisté à se donner dans un premier temps une définition syntaxique (à l'aide de propriétés formelles, donc) du complément de verbe, en tant qu'il s'oppose à l'ajout d'une part, au complément dit « de phrase » d'autre part ; nous avons procédé ici essentiellement à un travail de documentation mettant en jeu des phrases forgées par les auteurs consultés : Bonami (1999), Delaveau (2001), Dubois-Charlier (2001), etc. (donc en un

sens attestées, puisque produites par d'autres que nous-même). Puis, à partir d'une liste de verbes, mentionnés comme étant susceptibles de se construire avec la préposition *dans* (*cf.* l'index de Dugas et Manseau, 1996), nous avons cherché des attestations de ces différentes combinaisons dans des bases de données telles que *Frantext*, *Glossanet*, etc. Les verbes signalés par Dugas et Manseau (1996) n'étant pas tous l'objet d'une attestation, nous avons complété le corpus attesté par des phrases forgées. Enfin, sur le corpus de phrases attestées et forgées ainsi rassemblé, nous avons procédé au test des propriétés retenues pour distinguer entre complément de verbe et ajout, donc nous avons forgé un corpus de phrases (qui correspond aux résultats de l'application des critères).

3.3 Le recours à une base de données

La constitution d'une base de données pour rassembler son corpus[10] n'est pas une pratique naturelle en linguistique (entendue comme non spécialisée en TAL). Ainsi, nous mettons ici en évidence les apports de ce type de traitement et leurs avantages.

3.3.1. AVANTAGES GÉNÉRAUX, POUR LA RECHERCHE, DE LA CONSTRUCTION D'UNE BASE DE DONNÉES

a) L'intérêt pour le chercheur lui-même dans la gestion de son propre travail :

- la construction d'une base de données permet une perspective cumulative, donc de ne pas recommencer à constituer un corpus à chaque nouvelle recherche[11], et évolutive tant par sa structure (modulable) que par son contenu : on peut insérer ainsi des données à volonté (ajouter des informations sans cesse), qui peuvent être modifiées en fonction des usages. On peut ainsi l'améliorer, l'affiner pour finalement obtenir ce que l'on souhaite exactement.

- la souplesse : une base de données, une fois que sa structure est bien définie[12], est plus souple et plus puissante qu'une simple liste sur papier, dans *Word*, dans *Excel*… car elle permet notamment des mises à jour, constantes et en cascade, de données identiques mais enregistrées à différents endroits par exemple.

- le stockage et l'organisation des données : la base de données permet de stocker une quantité quasi illimitée d'informations (ce qui n'est pas négligeable quand on sait le nombre de manipulations que l'on effectue sur un corpus) et elle permet d'organiser des informations de façon significative : ainsi, on peut avoir sous les yeux toutes les données associées à un énoncé (sa source, ses analyses formelle et sémantique, les classes distributionnelles…). Elle contient donc le corpus avec des informations différentes : contextuelle, syntaxique, lexicale, sémantique… Elle permet en

quelque sorte de mieux voir les données (on peut proposer des vues sur les données, par exemple, le nombre d'enregistrements pour tel verbe…). Mais ce point de vue reste celui de la personne qui constitue la base de données. Ainsi ce qui compte pour établir une base de données, c'est de savoir ce que l'on souhaite en faire.

- le traitement automatique : la base de données permet aussi de récupérer des informations selon des critères de sélection (par exemple, on peut extraire la liste de tous les noms présents dans les SN introduits par *dans*, on peut aussi sélectionner tous les énoncés issus d'une même source (c'est ce que permet la table « Source », notamment si l'on veut faire une étude sur un journal particulier, sur un auteur particulier… Le traitement automatique des requêtes évite donc des manipulations fastidieuses à la main et offre un gain de temps qui permet d'approfondir la recherche et de mieux voir d'un coup d'œil les régularités. Il permet également d'opérer des analyses quantitatives (par les requêtes, les tris, les décomptes…), qui permettent alors de s'insérer davantage dans le courant de la linguistique de corpus. On peut, par exemple, s'interroger sur le type de nom qui est le plus fréquemment employé avec *dans* et les verbes de mouvement.

b) L'intérêt scientifique d'une circulation de la recherche et le fait qu'il existe relativement peu de corpus électroniques disponibles.

On peut diffuser l'information contenue dans une base de données : chaque linguiste, face à la spécificité de sa recherche, forge son propre corpus. Par la saisie d'un corpus dans une base de données, nous souhaitons rendre celui-ci accessible à d'autres linguistes pour plusieurs raisons : la première, c'est que nous nous sommes rendue compte, au fil de nos lectures (articles, revues… de linguistique), que nous n'avons pas accès aux corpus sur lesquels ces écrits ont été produits. Or cela nous aurait permis de vérifier les dires de certains linguistes, de compléter leur analyse sur le même corpus de base et de le compléter par de nouveaux énoncés pour confirmer, ou infirmer, ces dires. Ainsi, nous n'avons pas accès aux corpus analysés par Gross par exemple, or, il nous a semblé que certaines de ses analyses et conclusions n'étaient pas tout à fait exactes, mais seulement par rapport aux extraits de corpus qu'il nous donne et par rapport à notre propre corpus. La théorie n'est donc pas reproductible puisqu'on peut ne pas arriver aux mêmes conclusions. Chaque corpus construit par un linguiste meurt donc avec lui. Tant d'heures de recherche d'occurrences qui se perdent… L'accès au corpus des autres linguistes permettrait de gagner du temps et d'approfondir davantage la recherche. Ainsi, ce que nous avons recueilli pour la préposition *dans* peut servir à d'autres linguistes, leur permettre de mettre en évidence d'autres phénomènes que nous n'avons pas analysés (par exemple, quelqu'un qui travaille sur les temps grammaticaux pourra peut-

être y trouver des choses). De ce fait les informations contenues dans une base de données sont consultables et réutilisables par d'autres personnes.

De plus, il existe en France très peu de corpus électroniques disponibles[13], facilement accessibles (sur le français) qui puissent nous aider dans l'établissement de notre corpus d'étude. Il suffit pour s'en rendre compte de faire une recherche sur le Web avec le mot-clef « corpus » ou « base de données linguistique » ou « corpus linguistique » (les résultats sont probants !). Actuellement, le concordancier en ligne *GlossaNet*, le *TLFi*, le *Web*, le *Dictionnaire de l'Académie française*, *ABU* : la Bibliothèque Universelle, le site *Elicop* (Étude Linguistique de la Communication Parlée) sont disponibles et accessibles gratuitement alors que *Frantext*, *Le Monde Diplomatique*, *Le Petit Robert Multimédia* (ou autres corpus sur CD-Rom) restent sous le coût d'une licence (donc payants). Malheureusement, les corpus actuellement accessibles sont peu diversifiés (beaucoup sont centrés sur la littérature) ; ainsi, seul *GlossaNet* permet d'oublier pour un temps la recherche d'occurrences dans la presse munie d'un crayon !

3.3.2. AVANTAGES DE LA BASE DE DONNÉES LINGUISTIQUE ICI CONSTITUÉE : LE CORPUS EST CONSTITUÉ D'ÉNONCÉS MUNIS DE LEUR ANALYSE. Par l'informatisation de notre corpus, nous nous inscrivons dans le courant des linguistiques « de corpus » qui consiste en « l'utilisation de corpus annotés, de grande taille, variés et assortis d'outils d'exploration puissants, permettant d'observer plus finement les phénomènes » (Habert *et al.*, 1997). Par rapport aux faits, nous définirons notre corpus comme un regroupement de phrases isolées les unes des autres (absence de paragraphes, de textes…), mais ayant en commun l'usage de la préposition *dans*. Ces énoncés sont issus de sources différentes (presse, littérature… nous n'avons pas voulu distinguer des niveaux de langue différents et, par exemple, ne travailler que sur du « littéraire », ou que sur du « journalistique ») et récoltés de deux façons : la première reste traditionnelle – la lecture minutieuse armé d'un crayon pour relever ce qui nous semble pertinent. La seconde repose sur l'utilisation du concordancier *GlossaNet* : après avoir saisi nos requêtes de type [<dissoudre> dans] dans notre profil *GlossaNet* (notre recherche étant centrée sur la complémentation verbale en *dans*, nous souhaitions extraire des journaux une liste d'occurrences comportant les verbes se construisant avec cette préposition), le résultat de l'extraction nous était envoyé par courriel, il ne nous restait plus donc qu'à l'analyser et à saisir les occurrences dans notre base de données en suivant toujours la même procédure (les liens existants entre les tables, *cf*. Vaguer 2004) :

Etape 1 : Saisie dans la table « Source » de la provenance des énoncés récoltés (année, [auteur, titre], [journal, type de support : informatique, papier]) ;

Etape 2 : Saisie dans la table « Précisions sur la Source » pour spécifier l'article consulté, la page, le genre…

Etape 3 : La table « Identification distributionnelle du SP » contient l'énoncé retenu, ainsi que les propriétés syntaxiques du complément introduit par *dans* (les manipulations traditionnellement jugées pertinentes pour en permettre l'identification sont ici représentées : suppression, détachement, position préverbale, pronominalisation, test en *le faire*, entre autres) ;

Etape 4 : La table « Identification des constituants V, dét, N » permet de saisir chacun des constituants (en vue d'extraction automatique, par exemple, de l'ensemble des noms) et contient les conclusions de l'identification syntaxique du complément : est-il complément ou modifieur ?

Etape 5 : La table « Propriétés des Noms » permet une première analyse du nom en terme de classes d'objets, classe sémantique ou par ses propriétés morphologiques : est-il dérivé d'un verbe ?

À l'aide de cet échantillon d'emplois de la préposition *dans* (que nous souhaitons représentatif de l'ensemble de ses emplois en discours), nous avons pu mettre en évidence (Vaguer, 2004b) des régularités quant à l'utilisation de cette préposition, par le biais de manipulations réglées, et avancer ainsi dans son identification. Notre corpus comporte donc les énoncés de départ, mais aussi toutes les indications qui leur sont associées, tant du point de vue de la provenance de l'énoncé (source : auteur, genre, année, page…) que du point de vue de l'analyse de l'énoncé lui-même : son analyse syntaxique par le biais de manipulations (quel type de constituant, quelle structure de phrase, quelle fonction des constituants dans la phrase, etc.) et l'analyse de chacun de ses constituants (à quelle classe distributionnelle appartiennent-ils ?), son identité sémantique (locatif, approximatif…). À l'heure actuelle, notre base de données (nommée Zéphyr-V, V comme Verbe) rassemble 1 200 énoncés pourvus de leurs analyses syntaxique, lexicale et sémantique.

4. RÉFÉRENCES

Arrivé M.; Gadet F.; Galmiche M. 1986. *La grammaire d'aujourd'hui : guide alphabétique de linguistique française.* Paris : Flammarion.

Blanche-Benveniste, C. 2000. « Corpus de français parlé » *in* Bilger, M. (éd). *Corpus. Méthodologie et applications linguistiques*. Paris : Honoré Champion et PUP. (p. 15-25).

Bonami O. 1999. *Les constructions du verbe : le cas des groupes prépositionnels argumentaux*. Paris. Thèse de l'Université Paris VII.

Builles J.-M. 1998. *Manuel de linguistique descriptive. Le point de vue fonctionnaliste*. Paris : Nathan.

Chomsky N. 1969. *Structures syntaxiques*. Paris : Le Seuil.

Chomsky N. 1971. *Aspects de la théorie syntaxique*. Paris : Le Seuil.

Corbin P. 1980. « De la production des données en linguistique introspective ». *Théories linguistiques et traditions grammaticales*. Villeneuve-d'Asq : PU de Lille. (p. 121-179).
Delaveau A. 2001. *Syntaxe. La phrase et la subordination*. Armand Colin, Coll. Campus.
Dubois J.; Giacomo M.; Guespin L. 1999. *Dictionnaire de linguistique et des Sciences du langage*. Paris : Larousse (1ère éd. 1994).
Dubois-Charlier F. 2001. « Compléments de Verbe, de Proposition, de Phrase, d'Énoncé ». *Adverbe et Circonstant*. CLAIX. n°17. Aix-en-Provence : PUP. (p. 33-50).
Dugas A., Manseau H. 1996. *Les verbes logiques*. Montréal : Éditions Logiques.
Fillmore C. J. 1992. « "Corpus linguistics" or "Computer-aided armchair linguistics" » *in* Svartvik, J. (éd). *Directions in Corpus Linguistics*. number 65. Berlin : Mouton de Gruyter. (p. 35-59).
Gasiglia N. 2003. « Réflexions autour des coûts et bénéfices pour un linguiste qui recourt à des ressources électroniques et des outils informatiques dédiés à leur dépouillement : le cas d'une étude lexicale relative aux mots du football ». *Pré actes des 3èmes Journées de la linguistique de corpus*. Lorient (11-13/09/03). France.
Gleason H.-A. 1969. *Introduction à la linguistique*. Paris : Larousse.
Gross M. 1977. *Grammaire transformationnelle du français. Syntaxe du nom*. Paris : ASSTRIL.
Habert B.; Nazarenko A.; Salem A. 1997. *Les linguistiques de corpus*. Paris : Armand Colin / Masson.
Habert B. 2002. « Outiller les linguistes/outiller la linguistique : par où, par qui commencer ? ». Intervention à la table ronde *TAL et enseignement*. TALN'02 Nancy. 24/06/02.
http://www.limsi.fr/Individu/habert/Cours/PX/BHabertOutillerLaLinguistiqueTableRondeTALN02.pdf.
Leeman D. 2000. « Compléments circonstanciels ou appositions ? ». *Langue française*. n°125. Paris : Larousse. (p. 19-29).
Melis l. 2003. « Le groupe prépositif comme déterminant du nom » *in* Haderman, P., Van Slijcke, A., Berré, M. (éds). *La syntaxe raisonnée – Mélanges de linguistique générale et française offerts à Annie Boone*. Bruxelles/Paris : De Boeck/Duculot. (p. 235-250).
Mellet S. 2002. « Corpus et recherches linguistiques. Introduction ». *Corpus*. n°1. Nice : Publications de la Faculté des Lettres, Arts et Sciences humaines de Nice. (p. 5-12).
Milner J.-C. 1978. *De la syntaxe à l'interprétation. Quantités, insultes, exclamations*. Paris : Éditions du Seuil.
Normand C. 1972. « De quelques notions fondamentales (sur un enseignement d'initiation à la linguistique) ». *Langue française*. n°14. Paris : Larousse. (p. 32-56).
Picabia L., Zribi-Hertz A. 1981. *Découvrir la grammaire française. Une introduction active à la linguistique française et générale*. Paris : CEDIC.
Riegel M.; Pellat J.-C.; Rioul R. 1994. *Grammaire méthodique du français*. Paris : PUF.

Sampson J. 1994. « Susanne : a domesday book of english grammar » *in* Oostdijk, N., De Haan, P. (éds). *Corpus Based Research into Language*. Amsterdam : Rodopi. (p. 169-187).
Saussure F. de. 1972. *Cours de linguistique générale*. Paris : Payot. (1ère éd. 1916).
Sinclair J. 1996. *Preliminary recommendations on Corpus Typology*. Rapport Technique. EAGLES (Expert Advisory Group on Language Engineering Standards). CEE.
Vaguer C. 2000. *Il s'est trompé dans l'administration du médicament. Un ou des compléments de structure : « Dans + N_{action} » ? Naissance de la notion "complément d'apposition"*. Mémoire de DEA. Université de Paris X–Nanterre.
Vaguer C. 2004a. « Constitution d'une base de données : les emplois de *dans* marquant la "coïncidence" ». *Revue Française de Linguistique Appliquée*. IX-1. (p. 83-97).
Vaguer C. 2004b. *Les constructions verbales "V* dans *GN". Approches syntaxique, lexicale et sémantique*. Thèse de doctorat. Université de Paris X–Nanterre.
Vaguer C. 2005a. « Une base de données comme moyen de communication scientifique ? ». Actas-I, *IXème Simposio International de comunicación social*, organisé par le *Centro de lingüistica Aplicada* y *El Ministerio de Ciencia Tecnologia, y Medio ambiente*. Santiago de Cuba. (p. 134-138).
Vaguer C. 2005b. « De l'utilité d'un corpus en syntaxe, mais quel corpus ? ». *in* Vergely P (éd.). *Rôle et place des corpus en linguistique.* Actes du Colloque JETOU'2005. (p. 101-114).
Vandeloise C. 1986. *L'espace en français*. Paris : Le Seuil.

3. NOTES

(1) Tel que Dubois *et al.* (1999, p. 123) le définissent : « L'*univers* est l'ensemble des énoncés tenus dans une circonstance donnée, tant que le chercheur n'a pas décidé si ces énoncés entraient en totalité ou en partie dans la matière de sa recherche ».
(2) La représentativité est pour Gleason (1969, p. 158) un des problèmes essentiels liés à la constitution et à l'utilisation d'un corpus de matériaux, fournis par un ou plusieurs informateurs, et à partir duquel le linguiste doit écrire sa description de la langue. Le problème lié à la représentativité d'un « échantillon » de langue que forme le corpus, c'est que « certains traits grammaticaux ne sont pas fréquents ; ils risquent de ne pas être représentés de façon valable dans un corpus réuni au petit bonheur. D'autres traits, au contraire, sont très courants : même une quantité restreinte de matériaux suffit à les illustrer bien plus qu'il n'est nécessaire pour établir ou confirmer une analyse ». Normand (1972, p. 34) résumait ainsi les propos de Gleason : « des traits importants de la langue peuvent ne pas être représentés et des traits ordinaires l'être trop souvent ».
(3) « Un corpus ne peut être clos et exhaustif que dans le cadre d'une monographie… Il sera étudié en tant que tel, sans pouvoir prétendre à être représentatif d'autre chose que de lui-même ni à ouvrir sur aucune forme de généralisation ou modélisation » (Mellet 2002, p. 6).
(4) *Cf.* la caricature proposée par Fillmore (1992, p. 35): « Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact !", grabs his pencil, and writes something down. Then he paces around for new hours in the excitement of having come still closer to knowing what language is really like. (There isn't anybody exactly like this, but there are some approximations.) Corpus

linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations) ». L'idéal pour Fillmore serait que les deux types de linguistes soient réunis en un seul homme.

(5) « En français, le terme *informateur* peut prêter à confusion : il fait souvent penser à un indicateur, c'est-à-dire à quelqu'un qui fournit des renseignements à la police ou à un autre service plus ou moins officiel… En anglais, la confusion n'existe pas car il existe deux termes distincts : *informant* (celui qui fournit des renseignements à la police) et *informer* (celui qui fournit des renseignements à un journaliste, à un linguiste, etc.) » (Builles 1998, p. 60).

(6) L'emploi de corpus forgés permet au linguiste d'avoir « la langue accessible à travers une série toujours ouverte de nouveaux énoncés, spontanés ou provoqués… » (Riegel *et al.* 1994, p. 19). Et d'un point de vue quantitatif, le fait d'avoir accès à la langue dans son ensemble, et non uniquement à un échantillon (comme c'est le cas avec les corpus attestés), offre d'autres possibilités : « N'étant plus limités en nombre, les échantillons de performance étayent les hypothèses sur la langue, mais permettent aussi leurs vérifications en les confrontant à de nouvelles données » (*Ibidem*).

(7) Insistons sur le fait que rien n'est dit sur le recueil des données (comment on procède, sur quoi on opère, sur quels types de données). Finalement, la notion de « corpus » semble acquise et admise par l'ensemble des linguistes, qui l'emploient sans juger utile de la définir, comme allant de soi : la consultation de différents ouvrages (dont l'analyse est proposée ici) nous a permis d'observer que cette notion est souvent esquivée, ou non explicitée.

(8) « Moi, je suis de la France. Je ne dis pas : je suis la France. Je suis de la France. Toutes mes pensées, toutes mes façons d'être, toutes mes sensations, toutes mes vibrations, elles sont de la France » (Habert *et al.*, 1997, p. 9). Cet exemple extrait du corpus *Mitterand1* met bien en évidence qu'il n'est pas facile d'établir des distinctions tranchées entre les réalisations langagières jugées acceptables et celles jugées non-acceptables, puisque les constructions employées par F. Mitterand paraissent pour certaines agrammaticales. Or le Président les a employées et son insistance montre qu'il est conscient des structures énoncées (elles ne relèvent pas du lapsus).

(9) Comme cela se fait dans les recherches actuelles en linguistique de corpus : cf. Habert *et al.* (1997) et plus récemment les communications de Gasiglia, Arnaud, Alves, Fujimura, Manguin… aux 3èmes Journées de la Linguistique de Corpus (Lorient, septembre 2003).

(10) Nous entendons par *corpus*, une banque de données ouvertes qui sera alimentée et étoffée régulièrement en fonction des exemples rencontrés et des préoccupations de recherches. Notre corpus sera donc centré sur des énoncés constitués de la préposition *dans* et on le jugera saturé pour des raisons matérielles au moment de finaliser notre thèse.

(11) Il faut, pour ce faire, bien entendu travailler sur le même sujet.

(12) Notons, toutefois, qu'une base de données nécessite un travail long et fastidieux de mise en place : en effet, il faut, dans un premier temps, définir quels sont les éléments que l'on veut y voir figurer et comment on souhaite que cela s'organise (la mise en place de liens entre les tables n'est pas évidente). Mais cette formalisation permet d'avancer dans la compréhension du phénomène étudié puisqu'il faut, à ce moment-là, se demander ce qu'on cherche à mettre en évidence, ce qu'on veut voir apparaître, etc. Si des efforts sont investis dans la constitution d'une base de données, il y a ensuite un « retour sur investissement » (Habert, 2002) non négligeable.

(13) Or, tout linguiste travaillant sur un « corpus » (comme nous l'avons mentionné en première partie de cet article), il y a beaucoup de données riches qui restent inaccessibles.

TAL ET CORPUS

QU'EST-CE QU'UN « CORPUS HOMOGÈNE » ? RÉFLEXIONS À PARTIR D'EXPÉRIENCES EN EXTRACTION ET RECHERCHE D'INFORMATION

Patrice Enjalbert
Université de Caen − GREYC, CNRS

1. EXTRACTION D'INFORMATION ET CORPUS

Cette communication propose une réflexion méthodologique sur la constitution et l'étude de corpus pour un type spécifique d'applications TAL : *l'Extraction d'Information* (EI). Rappelons brièvement en quoi consiste cette « technologie » (Piacenza, 1997 ; Poibeau, 2003 ; Dupont *et al.*, 2002). Il s'agit d'extraire un ensemble d'informations factuelles, étroitement « ciblées », à partir d'un ensemble de textes relevant d'un domaine déterminé. Des exemples couramment donnés sont : opérations financières, extraites à partir de dépêches d'agences ; nouveaux procédés industriels relevés dans les revues spécialisées du domaine technologique en question ; suivi d'un patient dans des compte-rendus d'hospitalisation, etc. Les informations extraites peuvent être présentées à l'utilisateur ou entrées dans des bases de données. Les méthodes utilisées vont d'un enchaînement assez « classique » de procédures de traitements morphologiques, syntaxiques, sémantiques… à la recherche de motifs syntaxico-sémantiques prédéfinis, éventuellement acquis par apprentissage.

Cette tâche spécifique − popularisée et développée à travers la série de conférences MUC dans les années 1990 (Piacenza, 1997) − se combine aujourd'hui souvent à des systèmes de *recherche d'information* (RI), les techniques de l'EI permettant une analyse linguistique relativement « profonde » (en regard des analyses statistiques plus communes, et tout en restant bien évidemment légères et partielles), récupérant des « bribes d'information structurée », utilisables dans des procédures de recherche.

La référence à un corpus de travail est ici absolument essentielle. Il est hors de question d'obtenir des performances « suffisantes » sur du texte tout

venant. On voit en effet immédiatement que la notion de « domaine de connaissances » est importante, mais aussi quelque chose comme le « genre » : dépêches, articles, ouvrages didactiques… On invoque souvent l'idée d'une *homogénéité* du corpus (Poibeau, 2003), mais sans que cette notion très pragmatique ne soit vraiment définie et étudiée. Or, la délimitation d'un corpus adéquat est fondamentale pour déterminer la faisabilité de la tâche et traiter les problèmes de portage, que ce soit sur de nouveaux textes pour une tâche définie, ou pour réutiliser un ensemble de connaissances (et de procédures) linguistiques sur une nouvelle tâche. Les enjeux pratiques sont donc tout à fait significatifs.

La présente étude s'appuie sur un ensemble d'expériences, menées dans notre équipe, pour tenter de caractériser plus précisément ce que peut être un *corpus homogène* pour ce type de tâche. Les travaux en question seront rapidement présentés dans la section 2. On voit immédiatement se dessiner les deux facettes de la notion d'homogénéité : en termes d'information (et de domaine de connaissances sous-jacent) d'une part, et de formes linguistiques qui la « portent » de l'autre (section 3). La notion harissienne de Langue de Spécialité en fournit une première illustration « de référence », mais beaucoup trop restrictive : une analyse plus approfondie fait apparaître un riche déploiement de *formes linguistiques spécifiques* mises à profit dans la pratique en EI/RI (section 4 et 5). Nous pourrons alors tenter une synthèse et requalifier le principe d'homogénéité, sans doute trop « global », au profit d'une sorte de précepte opératoire prônant d'étudier une variété de structures linguistico-informationnelles stables dans un corpus donné (section 6).

2. EXPÉRIENCES

2.1. Extraction d'information

Cette technologie vise à rechercher − c'est-à-dire localiser et extraire − dans un corpus de textes très homogène thématiquement, un type très précis d'informations factuelles simples. Les informations extraites sont décrites dans des *fiches* (*templates* dans le jargon américain) qui peuvent alimenter une base de données, ou être restituées à l'utilisateur, par exemple sous forme de résumé. Les textes traités seront typiquement des dépêches d'agences (comme dans l'exemple de la figure 1) ou des articles de revues spécialisées, ou encore de certaines rubriques de presse généraliste. Un autre exemple, sur lequel nous travaillons, concerne un corpus de constats amiables d'accidents de la route (figure 2). Les informations extraites concernent le nombre d'impacts, les véhicules concernés, leurs conducteurs, les parties touchées, ou encore le lieu de l'accident (ville, route, carrefour, etc.).

On voit qu'il s'agit d'une forme de compréhension très limitée, partielle, sélective des textes. Les méthodes employées varient, mais on trouvera assez largement aujourd'hui les étapes suivantes :

- Repérage et analyse des groupes nominaux désignant des « entités » intéressantes pour la tâche fixée. Cette opération pourra être menée grâce à des grammaires locales (ou de manière équivalent des automates à états finis).
- Mise en relation des entités. Ici, on exploitera une structure verbes (ou déverbaux) − actants. Ceci peut être réalisé par la méthode « populaire » des *patrons d'extractions*, décrivant les enchaînements de groupes verbaux et nominaux possibles pour exprimer l'information attendue. Par exemple pour le texte de la figure 1, on aura un patron du type :

  <appointment> of <person> to <position>

  D'autres schémas pertinents seraient :

  <person> <retires> as <position>
  <person> <is succeeded> by <person>
  etc.

  Mais on peut préférer une analyse syntaxique plus classique, souvent aidée par un typage et autres informations sémantiques.
- Un certain nombre d'opérations de mise en cohérence, au niveau du *texte*, des « bribes d'information » ainsi extraites proposition par proposition doivent alors être réalisées. Typiquement, des liens de coréférence (nominale ou pronominale) doivent être établis.

Nous reviendrons sur ces différentes étapes dans les sections 4 et 5 pour en faire une analyse plus précise, en relation avec la question posée de l'homogénéité de corpus. Le lecteur intéressé pourra consulter un certain nombre d'ouvrages ou d'articles de synthèse tels que Piacenza (1997), Poibeau (2003) et Dupont *et al.* (2002).

Texte traité (dépêche)

CHARLOTTE, N.C., Sept. 13 ICNWI - United Dominion Industries Ltd. (NYSE, TSE : UDI), a manufacturer of diversified engineered products, today announced the appointment of John G. MacKay, 56, to the newly created position of executive vice president-Europe. Mr. MacKay, who has been president of United Dominion's Industrial Products Segment, will be responsible for working with various operating units and the corporate staff to enlarge the company's presence in Europe through acquisitions, joint ventures and other strategic alliances. He will be based in Europe at a site to be determined.

Résumé produit

United Dominion Industries Ltd. appoints John G. MacKay as executive vice president.

Fiches extraites (templates)

Trois "entités" ont ici été identifiées : un "événement" de succession, l'organisation et la personne concernée. Les champs de chaque fiche spécifient les liens avec les autres entités ou des informations conservées sous forme de texte, mais dont on a repéré et typé la nature (poste, nom, titre…).

<EVENEMENT-DE-SUCCESSION-c2O97.english.txt-16> :=
ORGANISATION: <ORGANISATION-c2O97.english.txt-43>
POSTE: " executive vice president"
PERSONNE: <PERSONNE-c2O97.eng1ish.txt-48>
RAISON-DU-POSTE-VACANT: AUTRE-INCONNU
<ORGANISATION-c2O97.english.txt-43> :=
NOM: "United Dominion Industries Ltd."
AUTRES-NOM: "United Dominion", "UDI"
TYPE: COMPAGNIE
<PERSONNE-c2097.eng1ish.txt-48> :=
NOM: "John G. MacKay"
AUTRES-NOM: "MacKay"
TITRE: "M."

Figure 1 : Extraction d'information. Mouvement de personnels dans la direction de société. D'après (Gaizauskas et al., 1997).

Étant arrêté momentanément sur la file de droite du Boulevard des Italiens j'avais mis mon clignotant j'étais à l'arrêt et m'apprêtant à changer de file. Le véhicule B arrivant sur ma gauche m'a serré de trop près et m'a abîme tout le côté avant gauche.

Figure 2 : Extrait d'un corpus de Constats d'accident (texte A7).

2.2. Recherche d'informations géographiques

Il est inutile d'insister sur l'importance aujourd'hui acquise par les technologies de la *Recherche d'Information* c'est-à-dire, dans la forme la plus courante, visant à rechercher et sélectionner, dans un fond documentaire, des documents pertinents sur un sujet donné. Les méthodes utilisées actuellement (majoritairement) ne font nullement appel à des méthodes linguistiques, mais utilisent plutôt des techniques numériques, statistiques ou autres, pour repérer les mots ou termes les plus significatifs du contenu, qui en constitueront une indexation. Toutefois, les limites de ces méthodes sont également bien connues de tout utilisateur, et une nouvelle tendance de recherche vise à développer des formes (limitées) d'analyse de contenu, en s'inspirant notamment de l'EI. On pourra aller jusqu'à envisager des formes particulièrement ambitieuses de RI, dans lesquelles le système fournirait non pas un document, pertinent par rapport à un problème que se pose l'utilisateur, mais directement la réponse. Ces nouvelles technologies ont pour nom *Question Answering* ou encore *Answer Extraction* dans le jargon du domaine (Voorhees, 2001).

L'application traitée ici ne va pas jusque-là, mais met néanmoins en œuvre des méthodes sémantiques pour améliorer les procédures de recherche d'information. Elle concerne un type particulier de document, le document géographique − aujourd'hui massivement produit et utilisé par toutes sortes d'administrations, publiques ou privées, chargées d'administrer ou gérer un territoire, et bien sûr par les géographes eux-mêmes. L'objectif très général du projet GéoSem[1] est alors de développer des outils permettant d'accéder le plus aisément, et de la manière la plus pertinente possible, à l'information contenue dans ce type de documents.

Or la spécificité de l'information géographique est de relier un *phénomène* observé (le *quoi*) à une *localisation géographique* (le *où*), ainsi que, très souvent, à une certaine *période temporelle* (le *quand*). Cette propriété s'observe immédiatement sur le texte comme on pourra s'en rendre compte dans la figure 3, et contribue très fortement à le *structurer*. Corrélativement, une *requête* naturelle de recherche documentaire portera sur un triple critère Phénomène-Espace-Temps : « où trouver des informations sur tel phénomène dans tel espace à telle période ? », l'une des composantes étant évidemment susceptible de faire défaut. Exemples de telles requêtes : « retard scolaire dans l'Ouest de la France dans les années 1950 », « évolution de la scolarisation entre 1960 et 1970 », ou encore : « politiques de sécurité maritime dans la manche ». Une réponse sera un ensemble de *passages* dans le corpus reliant ces trois critères.

De 1965 à 1985, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible ***dans le Sud-Ouest et le Massif central***, modérée ***en Bretagne et à Paris***, l'augmentation a été considérable ***dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne*** où les effectifs ont souvent plus que doublé.

Figure 3 : Extrait du corpus géographique

Un ensemble de méthodes sont mises en œuvre, dont nous retiendrons ici deux aspects. En premier lieu il convient de procéder à une *analyse sémantique des expressions spatiales et temporelles :* les critères de forme pure, déjà très rigides pour des requêtes thématiques classiques, ne sont ici à l'évidence d'aucune pertinence. En second lieu, il nous faut établir les relations existant dans le texte entre les trois « dimensions » de l'information (Phénomène-Espace-Temps), ce qui peut se faire parfois au niveau de la phrase, mais nécessite souvent une *analyse discursive* (locale). Ainsi dans l'extrait de la figure 3, le critère « de 1965 à 1985 » porte sur les deux phrases citées (et peut-être au-delà). Ces méthodes et les résultats obtenus sont décrits dans Bilhaut *et al.* (2003 a et b).

2.2.1. CORPUS HOMOGÈNES : ENTRE INFORMATION ET LANGUE. À partir de ce rapide examen de quelques applications et des méthodes mises en œuvre, une première réflexion peut s'opérer sur la question qui nous intéresse. Remarquons d'abord que la notion d'homogénéité est abordée en référence à *un type de tâche* : extraction et recherche d'information en ce qui nous concerne. Deux aspects, *deux facettes*, apparaissent immédiatement :

- Homogénéité de *l'information* « véhiculée » par les textes du corpus.

La définition même des informations à extraire, leur format, leur codage *symbolique*, est dépendante d'un domaine de connaissance spécifique. Restreindre drastiquement le domaine des connaissances nécessaires est une condition *sine qua non* de faisabilité.

- Homogénéité des *formes linguistiques*

La technique des « patrons syntaxico-sémantiques » illustre bien le fait que l'on s'attend à trouver l'information cherchée sous certaines formes, constituant un ensemble relativement « fermé ». Mais, nous le verrons, l'exploitation de régularités linguistiques va bien au-delà.

La notion harrissienne de *langue de spécialité* (Harris, 1991) − *LS* en abrégé, ou *sous-langage* − fournit un premier cas, que l'on pourrait dire « emblématique », d'homogénéité en ce sens. En effet, une LS se définit à la fois par un domaine de connaissances spécifique (en général : scientifique ou technique, par exemple telle branche de la biologie) et des structures

linguistiques fortes et régulières, en l'occurrence structures argumentales des prédicats verbaux, extensibles à d'autres relations syntaxiques (Assadi & Bourigault, 2000), « reflet » de la structure de l'information elle-même. De fait, certains corpus traités en EI peuvent être assez bien caractérisés comme relevant d'une LS : c'est le cas de certains textes scientifiques, mais aussi, pensons-nous, en sortant du cadre initial de « sous-langage d'une science », d'une certaine « littérature » de dépêches financières.

Néanmoins, si la notion de LS nous fournit un bon point de référence, elle ne rend pas compte de toutes les régularités observées et mises à profit en EI. Les corpus traités y sont en général nettement plus complexes. Ils ne concernent (souvent) pas une « science » ou une « technique » particulière, mais des domaines de connaissances moins fortement structurés, l'expression relevant plutôt d'une langue « commune » avec quelques spécificités plus techniques. Qui plus est, un même texte peut comporter des informations de nature assez diverse, au sein desquelles le système d'EI devra détecter, sélectionner, les informations pertinentes : pensons par exemple à des textes journalistiques. Par contre, l'examen révèle tout un ensemble de régularités, alliant information et formes d'expression, que nous allons maintenant examiner.

2.2.2. VARIATION 1 : « CLASSES D'EXPRESSIONS ». Une première remarque concerne la variété des structures syntagmatiques mises en jeu. Le domaine de l'information géographique fournit une bonne illustration de cette idée. La figure 3 a mis en évidence les trois types d'expressions caractéristiques : expressions spatiales (« dans le Sud-Ouest et le Massif central », « dans les départements de la grande banlieue parisienne »,…) ; expressions temporelles (« De 1965 à 1985 », d'autres exemples seraient : « dans les années 1950 », « au début des années 1950 »,…) ; et expressions évoquant le phénomène sociologique lui-même (variation des effectifs scolaires) : évidemment plus complexe mais dont on peut extraire, dans une perspective de recherche d'information, les groupes nominaux, significatifs de la thématique.

Les expressions temporelles sont particulièrement « rigides ». On peut les décrire avec une bonne approximation par des grammaires syntagmatiques locales, tout en calculant une représentation sémantico-référentielle comme un intervalle entre deux dates. Cette régularité s'étend d'ailleurs bien au-delà du corpus de « documents géographiques », dans des articles de presse géopolitique par exemple. Les expressions spatiales sont susceptibles d'une analyse et de traitements similaires. Leur structure générale combine des opérateurs de positionnement géométrique et de caractérisation sociologique, appliqués sur des régions désignées par des toponymes. Mais le lexique est évidemment d'un autre ordre de grandeur et d'une plus grande variabilité selon les textes. Enfin, dans la composante « phénomène » on trouve la

situation ordinaire de l'analyse thématique en recherche d'information et des structurations de type « terminologique » : champs lexicaux, structure tête-expansion, et ontologies associées.

Des remarques similaires peuvent être faites sur la tâche d'EI « standard ». Les expressions nominales pertinentes (décrivant les « entités » considérées comme intéressantes pour l'application visée) forment un ensemble relativement fermé. Par exemple : positions dans une société pour le corpus « mutations » et parties de véhicules, ou encore description des lieux, pour le corpus « accidents ». Nous avons par ailleurs à analyser des syntagmes propositionnels. Mais là encore, la limitation des informations extraites permet de délimiter ensemble de verbes (et déverbaux) et de constructions argumentales de ces verbes, relativement restreint − et descriptibles avec souvent une bonne approximation par les patrons d'extraction.

En résumé, plutôt qu'une LS unique, l'analyse fait donc apparaître un ensemble de *classes d'expressions*, possédant leur propre structuration syntaxique et sémantique. L'homogénéité plus ou moins grande d'un corpus par rapport à ces classes s'évalue à la fois en termes de *structure* (structures syntaxiques et opérations sémantiques associées), qui peut être relativement simple dès lors que ces expressions sont extraites de leur contexte ; et en termes de *lexique*, qui peut être riche et nécessite des méthodes de constitution et de structuration appropriées, mais en même temps délimité et relativement « fermé » par la classe d'expressions visée.

2.2.3. VARIATION 2 : AUTRES FORMES LINGUISTIQUES MISES EN JEU. Le second groupe de remarques concerne le fait que le seul type de structure visée par la notion de LS est de nature syntaxique (ou syntaxico-sémantique). Or d'autres formes de régularité apparaissent très clairement, notamment au niveau du discours (ou du texte). Ce point peut être illustré par plusieurs exemples − dont nous retiendrons ici quelques-uns issus de nos propres travaux.

- La notion de *cadres de discours* introduite par M. Charolles. Rappelons que l'auteur désigne sous ce terme des segments homogènes par rapport à un critère sémantique (par exemple une localisation spatiale ou temporelle), spécifié par une expression détachée en initiale de phrase, appelée *introducteur de cadre* (par exemple l'expression temporelle « De 1965 à 1985 » dans la figure 3). Les cadres spatiaux et temporels ont une prégnance manifeste dans les documents géographiques et doivent impérativement être exploités dans la tâche de recherche d'information (Bilhaut *et al.*, 2003 b).
- *L'anaphore*. L'anaphore temporelle dans un corpus géographique, si elle répond à l'évidence aux contraintes générales, présente un ensemble de spécificités qui peuvent être exploitées. Par exemple on trouvera très peu

d'anaphores pronominales, et certainement encore moins des phénomènes d'ambiguïté associés et qui compliquent tellement les procédures de résolution « en général ». L'anaphore nominale est par contre importante mais les possibilités de reprises (par exemple par hypo- ou hyperonymes) sont limitées. Des considérations similaires sont sans doute valides pour l'anaphore spatiale.

- *Macro structure discursive*. Un séquencement d'ensemble apparaît à l'étude des divers corpus. Dans les constats d'accidents nous avons une structure narrative en trois phases : description de la situation initiale, déroulement de l'accident, suites éventuelles. Ceci peut être exploité par exemple pour décrire le lieu de l'accident ou le mouvement des véhicules. Dans les textes géographiques, différents modes de présentation se succèdent et s'articulent : exposé du problème ou du phénomène observé, descriptions comparatives de type historique d'une part et spatiales de l'autre, éventuellement hiérarchiquement organisées, synthèses, etc. Des procédures de recherche d'information utilisant des critères spatiaux et temporels, ou encore des procédures de résumé automatique peuvent exploiter cette structure.

On devrait encore mentionner ici les questions de « mise en forme matérielle », indices précieux de l'organisation de l'information au sein du texte. Bref, les corpus traités présentent tout un ensemble de *traits caractéristiques*, bien au-delà du syntagme, facteurs d'autres *formes d'homogénéité*.

.2.4. SYNTHÈSE : QU'EST-CE QU'UN CORPUS HOMOGÈNE ? Nous pouvons maintenant revenir sur notre question titre et synthétiser les analyses précédentes autour de trois remarques.

1) La notion d'homogénéité doit s'entendre relativement à une *tâche* particulière, du type EI/RI pour ce qui nous concerne. D'autres pourraient sans doute appeler le même type d'analyse, autour de l'extraction de terminologie par exemple. On pourrait aussi imaginer un prolongement relatif à des faits de langues particuliers : le calcul de coréférence pourrait être un cas pertinent. Mais nous nous en tiendrons ici aux tâches précitées, objet de nos expériences.
2) L'approche de la langue qui se dessine dans ces travaux, selon nous, consiste en fait à observer et décrire, pour les besoins de la tâche, tout un ensemble de *microsystèmes*. Ceux-ci peuvent être de nature très diverses : locaux, comme les systèmes de localisation spatiale et temporelle géographique, ou globaux en tant que formes d'organisation du discours (coréférence, cadres…).
3) Ces microsystèmes intègrent indissolublement une structure particulière *d'information* et une *forme linguistique*. Tous les exemples des sections précédentes ont été analysés dans ce sens : expressions spatiales et

temporelles « géographiques » ou expressions nominales ciblées dans les premières phases de l'EI ; formes propositionnelles analysées pour l'EI, par exemple par la méthode des patrons d'extraction ; jusqu'aux cadres de discours qui, dans le document géographique, reprennent si étroitement, si fidèlement, la structure spécifique de *l'information* géographique.

Que dire dans ce cadre de la notion d'homogénéité ? En premier lieu (on pouvait s'en douter) que c'est, ainsi formulée, une notion beaucoup trop « globale ». Ce qui compte en pratique est la possibilité de décrire de manière uniforme toute une variété de « microsystèmes » (forme et sens), choisis en fonction d'un objectif donné. La notion d'homogénéité doit donc être conçue de manière *relative* à la classe de structures considérées, comme une bonne stabilité de cette classe particulière.

Conclusion décevante en un sens, puisque l'on est privé d'une caractérisation plus abstraite, plus intrinsèque, indépendante d'une tâche et de choix de description particuliers. Mais c'est un constat qui, à notre sens, ouvre (ou renforce) des perspectives pratiques intéressantes. Il signifie en effet que des progrès seront obtenus dans les tâches d'EI/RI par une description fine, précise, sur corpus, de ces « microsystèmes » − plutôt par exemple que par la découverte d'une « technologie miracle ». Une perspective, aujourd'hui de recherche, technologique demain, qui appelle la collaboration d'informaticiens et de linguistes − c'est-à-dire de spécialistes de l'information d'une part, et bons connaisseurs de la variété des formes linguistiques de l'autre.

3. RÉFÉRENCES

Assadi, H., Bourigault, D. 2000. "Analyses syntaxiques et statistiques pour la construction d'ontologies à partir de textes" *in* Charlet J. *et al.* (éd.) "Ingénierie des connaissances". Eyrolles. (p. 243-256).

Bilhaut, F., Charnois, T., Enjalbert, P., Mathet, Y. 2003. "Passage extraction in geographical documents". *Proc. Intelligent Information Systems 2003. New Trends in Intelligent Information Processing ans Web Mining*. Poland : Zakopane. 1-4 Juin 2003, (p. 121-130). (a)

Bilhaut, F., Ho-Dac, M., Borillo, A., Charnois, T., Enjalbert, P., Le Draoulec, A., Mathet, Y., Miguet, H., Pery-Woodley, M.-P., Sarda, L. 2003. "Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique". *Actes de TALN 2003*. Tome 1. (p. 315-320). (b)

Dupont, M., Vuillaume, J.-M., Victorri, B., Enjalbert, P., Mathet, Y. 2002. "Nouvelles tendances en Extraction d'Information". *Techniques et Sciences Informatiques*. Vol 21 n°1/2002. (p. 37-64).

Gaizauskas, R., Humphreys, K., Azzam, S., Wilks, Y. 1997. "Concepticons vs. Lexicons : An Architecture for Multilingual Information Extraction" *in* (Piacenza, 1997). (p. 28-43).

Habert, B., Nazarenko, A., Salem, A. 1997. *Les linguistiques de corpus*. Armand-Colin.

Harris, Z. 1991. *A theory of meaning and information. A mathematical Approach*. Clarendon Press.

Poibeau, T. 2003. *Extraction automatique d'information*. Hermès.

Piacenza, M.-T. (éd.) 1997. *Information Extraction*. Springer Verlag

Voorhees, E. 2001. "Overview of the TREC 2001 Question Answering Track". http://trec.nist.gov/pubs/trec10/t10_proceedings.html

4. NOTES

(1) Objet d'une collaboration entre le GREYC, le laboratoire ESO à Caen (géographie), l'ERSS à Toulouse (linguistique) et le groupe MEDIA (document électronique et technologies éducatives) de l'École Polytechnique Fédérale de Lausanne. Soutenu par le programme pluridisciplinaire « Société de l'information » du CNRS.

APPARIEMENT DE MOTS : PROPAGATION DES LIENS D'ÉQUIVALENCE À L'AIDE DE LA RELATION SYNTAXIQUE SUJET

Sylwia Ozdowska
ERSS, Université de Toulouse le Mirail

1. INTRODUCTION

L'appariement, c'est-à-dire la mise en correspondance, de mots ou expressions équivalente, dans des textes qui sont une traduction l'un de l'autre, représente un enjeu important, notamment pour ce qui est de la construction de ressources terminologiques multilingues. Divers systèmes visant à automatiser cette tâche ont vu le jour. Nombreux sont ceux qui privilégient l'utilisation de données statistiques dans leur stratégie de sélection du bon appariement (Daille *et al.* 1994 ; Gaussier 1995 ; Gaussier 1998), plus rares ceux qui exploitent des données linguistiques (Wu 2000 ; Hull 2001).

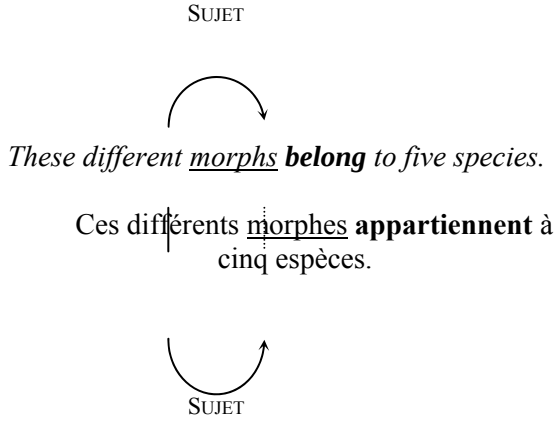
Dans cet article, nous proposons une méthode d'appariement de mots et de structures syntaxiques qui s'appuie principalement sur des connaissances linguistiques et, plus précisément, sur les relations de dépendance syntaxique identifiées, pour les deux langues source et cible, par les analyseurs syntaxiques de corpus SYNTEX. Le choix de cette méthode est guidé par les deux principaux objectifs que nous poursuivons : a) parvenir à un appariement précis à un niveau de granularité fin, c'est-à-dire celui des mots et des structures syntaxiques, et b) capter des appariements entre mots et structures syntaxiques peu fréquents et/ou spécifiques au corpus.

2. HYPOTHÈSE DE DÉPART

Nous reprenons à notre compte l'hypothèse formulée par Debili et Zribi (1996) selon laquelle *« les liaisons paradigmatiques peuvent aider à déterminer les relations syntagmatiques, et inversement »* et, plus particulièrement, l'idée que les relations de dépendance syntaxique sont

susceptibles, d'une part, de confirmer ou d'infirmer des liens d'appariement et, d'autre part, de créer de nouveaux liens. Le raisonnement est le suivant :

Si deux mots Ts_i et Tc_p sont appariés (*morphs* et *morphes* dans l'exemple) et s'il existe une relation de dépendance syntaxique entre Ts_i (*morphs*) et Ts_j (*belong*), d'une part, et entre Tc_p (*morphes*) et Tc_q (*appartiennent*), d'autre part, alors Ts_j (*belong*) et Tc_q (*appartiennent*) peuvent être appariés.



C'est ce mécanisme que nous nous proposons de mettre en œuvre, de manière automatique, dans notre étude. Et plutôt que de *« raisonnement par analogie »* (Debili et Zribi, 1996), nous préférons parler de propagation des liens d'appariement suivant les relations de dépendance syntaxique.

3. CORPUS

Le corpus de travail sur lequel s'appuie notre étude a été constitué dans le cadre d'une expérience, menée à l'INRA[1], sur l'enrichissement de la base de données terminologiques alimentée et exploitée par les traducteurs du service linguistique (Frérot *et al.*, 2001)[2]. Il s'agit par conséquent d'un corpus bilingue de traduction, avec le français pour langue source et l'anglais pour langue cible, qui de plus a été aligné automatiquement au niveau des phrases. Il compte environ 300 000 mots et comprend, par ordre décroissant d'importance, des articles de recherche, des articles de vulgarisation, un manuel d'utilisation de logiciel, des plaquettes de présentation, un contrat de licence, des résumés de monographies. Par ailleurs, il couvre différents domaines dont les plus représentatifs sont : l'agronomie, les sciences du sol, l'hydrobiologie, l'environnement, la biométrie et la modélisation, la génétique et l'amélioration des plantes, la pathologie végétale et la malherbologie. Ce corpus présente donc une certaine hétérogénéité tant du point de vue des thèmes abordés que des types de textes qui le composent.

4. OUTILS

Pour analyser le corpus d'étude, notre choix s'est porté sur les outils SYNTEX (Bourigault et Fabre, 2000). Il s'agit de deux analyseurs syntaxiques de corpus, l'un pour le français, l'autre pour l'anglais, qui prennent en entrée un corpus étiqueté et effectuent une analyse en dépendance de chaque phrase du corpus. SYNTEX prend en charge le repérage des sujets et objets des verbes (relations SUJET et OBJET)[3], avec une distinction au niveau de la relation lorsque le verbe est un verbe d'état (ATTRIBUT), le repérage des antécédents des pronoms relatifs (PROREL), ainsi que le rattachement des prépositions (PREP) et des compléments prépositionnels (PREP-D), celui des épithètes (ADJ), des adverbes ou encore des déterminants (DET).

L'analyse se fait de manière indépendante dans chacune des deux langues. Les traitements n'en restent pas moins homogènes étant donné que les deux analyseurs reposent sur un même principe de base, que les relations identifiées dans les deux langues sont les mêmes et leur représentation identique. Ce sont les principaux avantages de ces deux outils pour une étude comme la nôtre portant sur des corpus parallèles.

À partir des résultats de l'analyse syntaxique, SYNTEX extrait un ensemble de mots et de syntagmes qui servent de point de départ à l'ensemble du processus d'appariement que nous allons détailler dans les sections qui suivent.

5. PROCESSUS D'APPARIEMENT

Le processus d'appariement tel que nous le concevons comprend deux étapes. La première consiste à apparier les mots et syntagmes, dits candidats termes[4], sources (CTs) et cibles (CTc) extraits par les outils SYNTEX en se basant sur leur fréquence d'apparition dans des phrases alignées ; on parlera dans ce cas d'appariement global au niveau du corpus. La seconde, qui s'appuie sur les résultats obtenus à l'étape précédente, consiste à mettre en correspondance des CTs avec des CTc à un niveau local, c'est-à-dire phrase à phrase, à partir des relations de dépendance syntaxique identifiées par les mêmes outils. À terme, il s'agira de mettre en place un processus itératif où l'appariement global servira de base pour initier l'appariement local, les unités mises en correspondance de manière locale permettant à leur tour de poursuivre la propagation.

6. APPARIEMENT GLOBAL

6.1. Procédure d'appariement

L'appariement global met en œuvre une méthode utilisée dans de nombreux travaux (Gale et Church 1991 ; Kay et Röscheisen 1993 ; Ahrenberg *et al*.

2000). Il s'agit de comparer le nombre de fois où un CTs et un CTc apparaissent ensemble dans des phrases alignées, c'est la fréquence de cooccurrence, par rapport à la fréquence de chacun de ces CT sur l'ensemble du corpus. Soient freq(CTs), la fréquence du CTs, freq(CTc), la fréquence du CTc, freq(CTs, CTc), la fréquence de cooccurrence de CTs et CTc, la mesure d'association calculée, le Jaccard en l'occurrence, est la suivante :

$$j(CTs, CTc) = \frac{freq(CTs, CTc)}{freq(CTs) + freq(CTc) - freq(CTs, CTc)}$$

Pour cette expérience, nous avons choisi d'effectuer ce calcul uniquement pour les CTs et CTc dont la fréquence sur l'ensemble du corpus est supérieure ou égale à 5. De plus, pour un CTs donné, seuls les CTc pour lesquels $j(CTs, CTc) \geq 0,2$ sont retenus. L'ensemble de ces couples (CTs, CTc) forme ce que l'on appellera le lexique global.

6.2. Résultats et évaluation

Concernant l'appariement global, nous relevons tout d'abord les informations d'ordre général suivantes :

nombre de CTs avec freq(CTs) ≥ 5	3864
nombre de CTc avec freq(CTc) ≥ 5	3951
nombre de CTs appariés	3219
nombre moyen de CTc par CTs	3,29
valeur moyenne de j(CTs, CTc)	0,37

Tableau 1 : Appariement global

Nous avons évalué le lexique global avec les mesures habituelles de rappel et de précision :

Appel = nombre de couples (CTs, CTc) corrects/nombre de couples (CTs, CTc) de référence

Précision = nombre de couples (CTs, CTc) corrects/nombre de couples (CTs, CTc) proposés par le lexique global

Le lexique global offre un rappel de 80%. La précision est par contre très faible puisqu'elle n'atteint que 23,4%. Il est donc nécessaire de mener une réflexion concernant des techniques de filtrage du lexique global à mettre en œuvre, afin d'améliorer ce dernier résultat tout en gardant un rappel de même ordre.

7. APPARIEMENT LOCAL

7.1. Algorithme d'appariement

La première opération à effectuer avant de procéder à l'appariement local consiste à projeter le lexique global au niveau local, c'est-à-dire phrase à phrase. Pour ce faire on vérifie, pour chaque couple de phrases appariées, quels sont les mots appartenant à ces phrases qui ont été appariés au niveau global. Les couples de mots, issus de la projection et qui sont en relation d'équivalence au niveau des phrases alignées, constituent les couples amorces qui vont permettre d'initier le processus de propagation. Les résultats de la projection du lexique global aux phrases sont présentés dans le tableau 5. On obtient en sortie de la projection environ 40 000 couples amorces sur un ensemble de 7 055 couples de phrases alignées, ce qui donne en moyenne 6 couples amorces par phrase.

nombre de couples amorces	39852
nombre moyen de mots par phrase (anglais)	20
nombre moyen de mots par phrase (français)	22
nombre moyen de mots appariés par phrase	5,8

Tableau 2 : Projection du lexique global au niveau local

Afin de tester les performances d'un appariement local de mots par propagation des liens d'appariement le long des relations de dépendance syntaxique, nous avons choisi de travailler à partir de la relation SUJET et, plus précisément, à partir des noms qui sont régis par cette relation. Il en résulte que, dans le cas de relation SUJET, le sens de la propagation des appariements est celui qui va du régi, le nom, vers le recteur, le verbe.

Soient un nom anglais, Ns, et un nom français, Nc, formant le couple amorce (Ns, Nc)[5]. Soient Vs un verbe anglais et Vc un verbe français. Nous avons implémenté les deux règles d'appariement suivantes :

(1) si Ns est en relation SUJET avec le verbe Vs et Nc est en relation SUJET avec le verbe Vc, on apparie Vs et Vc ;
(2) si Ns est en relation Sujet avec le verbe Vs et Nc est en relation Objet avec le verbe Vc et si Vs est à la forme passive, on apparie Vs et Vc.

Ces deux règles d'appariement local par projection permettent donc de traiter deux types de cas :

(1) Ceux où les deux noms source et cible sont régis par la relation SUJET et où l'appariement des verbes se fait indépendamment de leur forme, active ou passive :

The fish ***are*** *generally* ***caught*** *when they migrate from their feeding areas towards their spawning grounds.*
Généralement les poissons ***sont capturés*** *lorsqu'ils migrent de leur zone d'engraissement vers celles de reproduction.*

(2) Ceux où l'un, le nom source, est régi par la relation SUJET et l'autre, le nom cible, est régi par la relation OBJET et où l'appariement se fait en fonction de la forme du verbe source, avec la condition que ce dernier soit à la forme passive :

The predictor ***can*** *then* ***be constructed****.*
On ***peut construire*** *le prédicteur.*
Activated sludge ***is submitted*** *to anaerobic conditions.*
En ***soumettant*** *les boues à des conditions d'anaérobie…*
The value of the sawn products ***must be maximised****.*
Il ***faut maximiser*** *la valeur des produits sciés.*
Activities in the field of human nutrition ***have*** *greatly* ***developed****.*
Il ***s'y est développé*** *une activité importante dans le domaine de la nutrition humaine.*

L'appariement par propagation conduit à établir des liens d'équivalence entre mots, comme c'est le cas de *catch*/*capturer* [6], *construct*/*construire*, mais aussi entre structures syntaxiques. En effet, à partir de l'appariement global *fish*/*poisson*, d'une part, et de l'appariement local *catch*/*capturer*, d'autre part, il est possible d'apparier les structures syntaxiques *catch_fish*/*capturer_poisson*. Nous parlerons donc, en ce qui concerne l'appariement par propagation, d'appariement syntaxique, premièrement en raison de l'utilisation des relations de dépendance syntaxique au cœur même du processus d'appariement et, deuxièmement, en raison du type d'unités concernées par l'appariement qui peuvent être aussi bien des mots simples que des structures syntaxiques.

7.2. Résultats et évaluation

La propagation des liens d'appariement basée sur la relation syntaxique Sujet/Objet a donné lieu à l'appariement de 1 591 couples de verbes. Nous avons validé manuellement 649 cas afin de constituer une base de référence pour l'évaluation de l'appariement local. Ont été évalués comme corrects aussi bien des appariements où un verbe simple en langue source correspond à un verbe simple en langue cible, que des appariements où un verbe simple en langue source correspond à une locution verbale en langue cible, et inversement, l'appariement ne concernant que le seul noyau verbal, soit la tête, de la locution :

[improve]
[conduire] à l'amélioration
[have] an influence
[influer]

Nous considérons en effet que, même s'il n'y a pas d'équivalence du point de vue du sens entre les verbes ainsi appariés, il n'en reste pas moins que l'appariement est juste du point de vue de la propagation. Par ailleurs, une fois trouvé l'appariement entre un verbe simple et la tête de la locution verbale qui lui correspond dans l'autre langue, nous pensons qu'il est possible de mettre en place des heuristiques permettant de retrouver l'appariement complet entre verbe simple et locution verbale.

Sur les 649 cas validés, 54 correspondent à des échecs, ce qui représente un taux de précision de 91,70%. Il est à noter que sur les 54 échecs relevés, 43 ont pour origine une erreur préalable à l'algorithme de propagation telle que : 1) une erreur dans l'analyse syntaxique, 2) une erreur au niveau du repérage du couple amorce ou 3) une erreur dans l'alignement des phrases.

1) *The drift* ***observed*** *in cyprinid alevins may also be density dependent.*
La dérive observée chez les alevins de cyprinidés ***peut*** *aussi* ***dépendre*** *de phénomènes surdensitaires.*

Dans l'exemple 1), le mot *observed* a été identifié par l'étiqueteur morpho-syntaxique comme verbe et non pas comme participe passé. L'analyse syntaxique a donc conduit à reconnaître *drift* comme sujet de *observed*. À partir du couple amorce (*drift*, *dérive*), le lien d'équivalence a donc été à tort propagé suivant la relation SUJET à *observe*/*dépendre*.

2) *The conditions of residual normality are examined using the Shapiro Wilk test, the other hypotheses* ***were checked****…*
L'hypothèse$_1$ de normalité ***est testée*** *par le test de Shapiro-Wilk, les autres hypothèses$_2$ sont vérifiées…*

Dans l'exemple 2), c'est le couple (hypothesis, hypothèse1) qui a été reconnu par erreur comme couple amorce, au détriment des couples (condition, hypothèse1) et (hypothesis, hypothèse2). En effet, le lexique global propose pour hypothesis un seul équivalent possible : hypothèse. Dès que l'algorithme de projection a repéré la présence de hypothesis et hypothèse1, il sélectionne (hypothesis, hypothèse1) comme couple amorce, sans vérifier s'il y a ou non d'autre possibilités, ici (hypothesis, hypothèse2). Il s'agit là d'une limite de l'algorithme de propagation qui demande à être amélioré.

Seuls 11 des échecs relevés viennent d'une erreur de l'algorithme de propagation même, ces échecs ayant pour origine la présence de reformulations qui interviennent lors du processus de traduction et dont nous présentons ci-dessous quelques exemples.

The radio transmitter ***is composed*** *of an electronic part and a battery and emits a pulsating signal.*

L'émetteur radio, composé d'une partie électronique et d'une pile, ***émet*** *un signal pulsé.*
The Jouy Research Center ***employs*** *11% of all INRA staff and, as in the past, will continue to play a major role…*
Rassemblant plus de 11% des moyens humains de l'Institut, le Centre de Recherche de Jouy ***est appelé*** *à jouer, comme par le passé, un rôle majeur…*
The multiplicative model sometimes ***contained*** *only three testers […] when it was run for the elementary diallels.*
Le modèle multiplicatif ***a*** *parfois* ***été appliqué*** *sur des diallèles élémentaires n'incluant que trois testeurs […].*

Ces cas de reformulation sont des cas de non correspondance que l'algorithme de propagation n'est pour le moment pas en mesure de repérer. En effet, la relation qui sert de base à la propagation est bien présente dans les deux langues ; *model* et *modèle* sont chacun sujet d'un verbe, il n'en reste pas moins que les verbes en question ne sont pas en relation d'équivalence et ne doivent donc pas être appariés.

Il existe par contre des cas de non correspondance qu'il est facile de repérer automatiquement. Il s'agit de ceux où la relation à l'aide de laquelle on souhaite propager les liens d'équivalence, la relation SUJET ou OBJET dans cette étude, est présente dans l'une des deux langues et pas dans l'autre.

7.3. Cas de non correspondance

Le tableau 3 (ci-dessous) présente la manière dont se répartissent les cas de non correspondance en termes de relations syntaxiques identifiées[7]. Il montre que l'on trouve la même relation dans les deux langues dans près de 65% des cas[8]. On dira alors qu'il y a un parallélisme dans les deux langues. Par conséquent, le parallélisme n'est pas respecté dans les 35% des cas restants.

Ces cas de non correspondance sont très intéressants à observer, dans la mesure où ils peuvent révéler des régularités interlingues et permettre d'identifier les relations susceptibles de servir de base à une propagation indirecte, c'est-à-dire à partir de relations syntaxiques qui sont différentes dans les deux langues. L'étude des cas de non correspondance en français, lorsque la relation est SUJET en anglais, nous a ainsi permis de mettre au jour un certain nombre de régularités que nous comptons exploiter afin d'étendre l'algorithme de propagation.

Ns en relation SUJET	2474	100%
Cas de correspondance (traités par l'algorithme de propagation)		
Nc en relation SUJET	1512	61,1%
Nc en relation OBJET (Vs au passif)	79	3,2%
Cas de non correspondance identifiés automatiquement		
Nc en relation PREP-D	419	17%
Nc sans relation syntaxique	243	9,8%
Nc en relation ATTRIBUT	125	5%
Nc en relation OBJET (Vs actif)	52	2,1%
Nc en relation PROREL	30	1,2%
Nc en relation EPI	21	0,8%

Tableau 3 : Répartition des cas de non correspondance

Relation Prep-d. Lorsque le Nc est relation PREP-D avec un autre mot de la phrase, autrement dit lorsqu'il est complément de nom, nous pouvons avoir affaire à :

a) une nominalisation ou, plus généralement, selon la terminologie de Vinay et Darbelnet (1958), une transposition verbe-nom :

> *The manner in which the fungus develops in its host is fairly well known.*
> *Le mode de développement du champignon dans son hôte est assez bien connu.*

b) une différence d'explicitation, qui est dans ce cas plus importante en français qu'en anglais :

> *A research team in Quebec has shown that the amount of herbicide residue in drainage water is linked to annual rainfall levels.*
> *Les travaux d'une équipe québécoise montrent que la quantité des résidus herbicides dans les eaux de drainage est liée au taux de pluies annuelles.*

c) une modulation (Vinay et Darbelnet, 1958)

> *In order to be able to visualise the inside of the logs, the researchers used a medical scanner.*
> *Pour visualiser l'intérieur des billons de bois, un scanner médical a été utilisé par les chercheurs.*

Relation Attribut. Cette relation, qui lie un nom à son attribut, peut indiquer la présence en français d'une locution verbale composée d'un verbe

d'état suivi d'un adjectif. Cette locution est à mettre en correspondance avec un verbe simple en anglais.

The intensity of diffraction reflections varies with the orientation of the sample.
On sait que l'intensité des raies de diffraction est variable suivant l'orientation de l'échantillon.

Relation Objet. La présence de cette relation peut être liée à une opposition construction personnelle/construction impersonnelle.

It seems that the allis shad has never completely disappeared from the Rhine and that a small population still exists today.
Bartl et Troschel in Roche semblent considérer que la grande Alose n'a jamais complètement disparu du Rhin et qu'il existerait actuellement une petite population.

Relation Prorel. Elle indique qu'en français le lien entre le nom et le verbe ne se fait pas de manière directe, mais par le biais d'un pronom relatif.

The learned society comprises about 400 members and meets every three years.
Cette société scientifique qui regroupe environ 400 membres, tient un congrès tous les 3 ans.

8. DISCUSSION ET PERSPECTIVES

Avec un taux de précision de 91,7%, l'appariement local par propagation offre un résultat comparable, voire supérieur, à ceux obtenus par Daille *et al.* (1994), précision variant entre 70 et 80% selon le nombre d'appariements validés, ou par Gaussier (1998), qui fait état d'une précision allant de 90 à 98% suivant le nombre de meilleures associations prises en compte. Wu (2000), quant à lui, estime ce taux à 81,5% et enfin Hull (2001) qui, privilégiant le taux de rappel, se contente d'une précision ne dépassant pas les 56%. Il reste à vérifier si on arrive à des résultats similaires en propageant les liens d'appariement à l'aide de relations syntaxiques autres que celle de SUJET/OBJET. C'est là une première piste de recherche que nous nous proposons de suivre. Parallèlement, nous menons une réflexion concernant les techniques de filtrage du lexique global et leur impact sur les taux de rappel et de précision de l'appariement global mais aussi sur ceux de la projection et de la propagation.

Par ailleurs, l'analyse des cas où le principe de propagation est mis en défaut est extrêmement enrichissante, à la fois du point de vue linguistique car elle permet de mettre au jour des régularités liées à la variation interlingue, et du point de vue de l'implémentation car elle offre la possibilité d'étendre l'algorithme de propagation afin de réduire le silence.

Enfin, notre réflexion porte également sur l'utilisation éventuelle de ressources exogènes, notamment de dictionnaires électroniques, susceptibles de compléter le lexique construit lors de la phase d'appariement global.

9. REFERENCES

Ahrenberg L.; Andersson M.; Merkel M. 2000. "A knowledge-lite approach to word alignment", *in* : J. Véronis (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*. Dordrecht : Kluwer Academic Publishers, p. 97-138.

Bourigault D.; Fabre C. 2000. "Approche linguistique pour l'analyse syntaxique de corpus", *Cahiers de Grammaire,* 25, Université Toulouse le Mirail, p. 131-151.

Brown P.; Della Pietra, S.; Mercer, R. 1993. "The mathematics of statistical machine translation : parameter estimation", *Computational Linguistics*, 19(2), p. 263-311.

Chuquet H.; Paillard M. 1989. *Approche linguistique des problèmes de traduction anglais/français*. Ophrys.

Daille B.; Gaussier E.; Langé J-M. 1994. "Towards Automatic Extraction of Monolingual and Bilingual Terminology", *Proceedings of the International Conference on Computational Linguistics (COLING'94)*, p. 515-521.

Debili F.; Zribi, A. 1996. "Les dépendances syntaxiques au service de l'appariement des mots. " *Actes du 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*.

Debili F. 1997. "L'appariement : quels problèmes ?" *Actes des 1ères JST 1997 FRANCIL de L'AUPELF-UREF*.

Déjean H.; Gaussier E. 2002. "Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables", *Lexicometrica*, numéro spécial *Alignement lexical dans les corpus multilingues*.

Gale W. A.; Church K. W. 1991. "Identifying Word Correspondences in Parallel Text", *Proceedings of the DARPA Workshop on Speech and Natural Language*.

Gale W. A.; Church K. W. 1993. "A program for aligning sentences in bilingual corpora", *Computational Linguistics*, 19(3), p. 75-102.

Guillemin-Flescher J. 1981. *Syntaxe comparée du français et de l'anglais*. Ophrys.

Frérot C.; Rigou C.; Lacombe A. 2001. "Approche phraséologique d'une extraction automatique de terminologie dans un corpus scientifique bilingue aligné". *Actes des 4èmes rencontres Terminologie et Intelligence Artificielle*. Nancy, p. 180-188.

Gaussier E. 1995. *Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues*, Thèse de doctorat, Paris VII.

Gaussier E. 1998. "Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora", *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics* (COLING/ACL'98), p. 444-450.

Gaussier E. 2001. "General considerations on bilingual terminology extraction", *in* : D. Bourigault, Ch. Jacquemin, M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology*, John Benjamins, p. 167-183.

Gaussier E.; Hull D. A.; Aït-Mokhtar S. 2000. "Term alignment in use", *in* : J. Véronis (éd.), *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht : Kluwer Academic Publishers, p. 253-274.

Hull D. A. 2001. "Software tools to support the construction of bilingual terminology lexicons", *in* : D. Bourigault, Ch. Jacquemin, M.-C. L'Homme (éds.), *Recent Advances in Computational Terminology*, John Benjamins, p. 225-244.

Kay M.; Röscheisen M. 1993. "Text-Translation Alignment", *Computational Linguistics*, 19(1), p. 121-142.

Véronis J. 2000. "Alignement de corpus multilingues", J.-M. Pierrel (éd.), *Ingénierie des langues*. Paris : Éditions Hermès, p. 115-150.

Véronis J. (ed). 2000. *Parallel Text Processing : Alignment and Use of Parallel Corpora*. Dordrecht : Kluwer Academic Publishers.

Vinay J-P.; Darbelnet J. 1958. *Stylistique comparée du français et de l'anglais*. Paris : Didier.

Wu D. 2000. "Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars", *in* : J. Véronis (éd.), *Parallel Text Processing : Alignment and Use of Translation Corpora*. Dordrecht : Kluwer Academic Publishers, p. 139-167.

10. NOTES

(1) Institut National de la Recherche Agronomique.

(2) Nous remercions A. Lacombe de nous avoir autorisée à utiliser ce corpus à des fins de recherche.

(3) Nous indiquons entre parenthèses les noms que nous utiliserons pour faire référence à ces différentes relations dans la suite de l'article.

(4) Mots ou syntagmes susceptibles d'acquérir le statut de termes spécifiques à un domaine donné.

(5) Les couples amorces sont en caractères soulignés dans les exemples, et les mots appariés par propagation en gras.

(6) Les mots et structures syntaxiques appariés par propagation sont présentés sous leur forme lemmatisée.

(7) Les noms des types de relation sont définis en 4.

(8) Si l'on admet que la relation SUJET, lorsqu'elle concerne un verbe au passif, peut être assimilée à la relation OBJET si cette dernière concerne un verbe à l'actif.

REPÉRAGE DES NON-PHRASES THÉMATIQUES GRÂCE À LA PLATE-FORME CONTEXTO(1)

Sylvie Porhiel
LaLLIC (Langage, Logique, Informatique, Cognition et Communication)
UMR 8139 CNRS, Université Paris IV, Paris.

Résumé
Cet article détaille les critères pris en compte pour repérer des non-phrases thématiques dans la plate-forme ContextO, c'est-à-dire des phrases sans verbe qui jouent un rôle dans l'organisation du discours en désignant les circonstances dans lesquelles il faut envisager un certain état ou une série d'événements. Une telle analyse a pour but d'extraire les non-phrases thématiques dans tout type de texte et, ce faisant, d'aider l'utilisateur à repérer la structure thématique d'un texte.
Notre exposé se déroule ainsi : nous commençons par situer le cadre de notre recherche dans son projet d'accueil. Nous exposons ensuite les données concernant les non-phrases sur le plan linguistique et sur le plan informatique, et proposons une analyse descriptive plus approfondie. La dernière partie explique comment les résultats de l'analyse linguistique ont été implémentés dans la plate-forme ContextO, les problèmes que cette implémentation a soulevés, et expose enfin les résultats concernant l'extraction automatique des non-phrases thématiques.

1. CADRE DE LA RECHERCHE ET PROJET D'ACCUEIL

Notre projet de recherche sur les introducteurs de cadre thématique dans des non-phrases (*non sentence* de Hollerbach (1994) ; c'est-à-dire des phrases sans verbe) s'inscrit dans un projet plus vaste, concernant le repérage de la segmentation thématique dans les textes au moyen de la plate-forme ContextO, mise au point par le laboratoire LaLICC (Langage, Logique, Informatique, Cognition et Communication) dirigé par J.-P. Desclés.

Dans la perspective textuelle de Charolles (1997), les introducteurs thématiques contribuent à partitionner l'information dans des blocs sémantiquement homogènes, en désignant les circonstances dans lesquelles il faut envisager un certain état ou une série d'événements. Selon la liste indicative proposée dans Porhiel (2004), les unités lexicales qui composent

cette classe discursive sont des prépositions (*au sujet de, à propos de, pour ce qui est de, concernant*, etc.) ou des adverbes – des anaphores résomptives – (*à ce sujet, à ce propos*, etc.). Ces unités participent à la cohésion textuelle : elles instaurent des liens cohésifs (Hasan et Halliday, 1978) inférables à partir d'indices linguistiques.

Pour repérer les introducteurs thématiques, nous utilisons la méthode d'exploration contextuelle (Desclés *et al.,* 1997 ; Minel *et al.,* 2001) qui identifie des connaissances linguistiques en les restituant dans leurs contextes et en les organisant en fonction de tâches spécialisées. Une règle d'exploration contextuelle définit un espace de recherche : il s'agit d'un segment textuel toujours déterminé à partir de la présence d'un marqueur déclencheur, ici les introducteurs thématiques ; des indices complémentaires doivent en général être recherchés dans l'espace de recherche en vue de confirmer ou d'infirmer la valeur sémantique de l'indicateur repéré (la position du déclencheur dans la phrase, par exemple). L'ensemble de ces connaissances linguistiques, supposées indépendantes du domaine dont traite le texte, sont organisées dans un modèle conceptuel (Ben Hazez et Minel, 2000 ; Minel *et al*., 2001) et exploitées par la plate-forme logicielle ContextO. Cette capitalisation des données linguistiques fournit à l'utilisateur un outil d'aide informatisé qui répond à ses besoins de fouille de texte (Minel, 2003), en l'occurrence le repérage, sur un plan généralement local, de la structure thématique d'un texte.

Dans ce cadre, les analyses sur les introducteurs thématiques ont principalement concerné leur fonctionnement dans le cadre de la phrase (Ferret *et al*., 2001 ; Porhiel, 2001b). Elles montrent que les introducteurs thématiques possèdent des caractéristiques linguistiques propres : **1)** morphologiques : certains marqueurs varient en nombre (*au chapitre de, aux chapitres de*), d'autres varient en temps (*en ce qui concerne, en ce qui concernait*), d'autres encore ont une forme résomptive (*au sujet de, à ce sujet*) ; **2)** syntaxiques : un introducteur thématique préfixe, prototypiquement, au moins une proposition et au plus un paragraphe. La différence entre une expression introductrice de cadre et une conjonction se fait en termes de dépendance, ce qui se traduit ici en terme positionnel. La préposition d'une phrase thématique se trouve en position initiale : ***Pour ce qui concerne*** *la France, les organisateurs de la conférence de Stockholm se montrent extrêmement critiques à l'égard du Minitel (...)* (AJ), alors que celle d'une phrase a-thématique dépend d'un constituant morphosyntaxique : *(...), les Italiens arrivent unanimement premiers* ***pour ce qui est de*** *se pomponner, (...)* (AJ). Certains groupes de mots (des adverbiaux) peuvent se placer avant les introducteurs thématiques ; tout comme ces derniers, ils sont en position détachées : *Mais* ***en ce qui a trait à*** *l'accès à cette technologie, l'écart entre le Nord et le Sud est encore plus grand que pour la répartition des autres ressources*. (AJ) ; **3)** ponctuationnelles : l'introducteur thématique

suivi de son complément est prototypiquement suivi d'une virgule (*cf*. exemples ci-dessus). Cette caractéristique ponctuationnelle constitue un critère robuste pour distinguer une phrase thématique d'une phrase a-thématique. La virgule peut toutefois ne pas apparaître dans le cas de patrons syntaxiques récurrents comme dans : *Mais c'est* ***en matière de*** *soutien public à la science et à l'innovation que l'avenir a été le plus lourdement hypothéqué* (AJ) ; **4)** sémantiques : les introducteurs thématiques instaurent des relations non spatio-temporelles. Ils sont de nature abstraite et acceptent, hors contexte, en termes de compatibilités lexicale et sémantique, tout type de complément. Toutefois, comme les prépositions n'instaurent pas de relation unilatérale (Mann et Thompson, 1988 ; Knott et Sanders, 1998), une même préposition peut potentiellement être un introducteur de cadre (*Au niveau du développement, (...)*) ou un complément circonstanciel (*Au niveau des épaules, (...)*). Il est donc indispensable d'inhiber les possibilités relationnelles non abstraites des prépositions, telles les relations spatiales ou avec une partie du corps, avec la préposition *au niveau de*. Ces caractéristiques ont été capitalisées dans la base de données de la plate-forme ContexO et ont été utilisées dans des règles déclaratives de repérage écrites en JAVA et actuellement en cours de test. En revanche, les études concernant les non-phrases en général, sont plus rares tant sur le plan linguistique que sur le plan informatique.

2. TRAVAUX EN RELATION AVEC LE SUJET

2.1. Travaux linguistiques

Sur le plan linguistique, les non-phrases sont des phrases orthographiques (ou textuelles). Ce sont des phrases « incomplètes » car elles n'ont pas de verbe et ne correspondent pas au patron syntaxique Sujet Verbe Objet du système linguistique français. Elles peuvent se subdiviser en :

- ellipses, au sens syntaxique, les non-phrases étant interprétées à la lumière d'une phrase précédente (Quirk et al. 1985 ; Riegel, Pellat, Rioul 1994) :

1) *Joseph se retourna*.
– *quelle réponse* ?
– ***au sujet de*** *la chambre*. (Frantext)

- fragments (Fries, 1952 ; Quirk *et al*., 1985 ; Nunberg, 1990 ; Riegel *et al*., 1994 utilisent le terme de *segment*). Ils résultent d'une utilisation particulière des signes de ponctuation dans le but de mettre certains éléments en relief ou de rendre le discours plus vivant :

2) *Il ne t'en a jamais parlé* ?

- En détail, non. Attends. Il me semble qu'il m'a raconté une anecdote, une histoire très jolie, trop. ***À propos de*** *Staline. Si c'est bien lui qui me l'a racontée…* (Frantext)

- non-phrases thématiques qui, pour leur interprétation, ne dépendent pas uniquement d'un cotexte linguistique (syntaxe et sémantique) et jouent un rôle dans le discours. Les non-phrases thématiques spécifient comment considérer une série d'événements. Elles ne se trouvent pas systématiquement en continuation directe avec la phrase précédente :

3) En réalité, les Australiens de souche sont inquiets pour leur avenir. ***Sur le plan de*** *la natalité d'abord : dès l'an 2000, leur population va baisser, (...).* ***Sur le plan des*** *valeurs ensuite : (...) les Australiens sont devenus blasés* (…). (AJ)

Dans les exemples 1 à 3, les non-phrases (i) sont toutes introduites par une préposition qui peut potentiellement être un introducteur thématique, (ii) sont construites sur le même patron syntaxique : Préposition+(Déterminant)+Complément [auquel on trouve parfois adjoint un Modifieur]. Aucune de ces non-phrases n'est analysable hors contexte et demande, pour son interprétation, au moins une phrase adjacente. Les relations interphrastiques ont été soulignées à maintes reprises (Dea, Belkin, 1978 ; Winter, 1978, *inter alia*) et sont d'autant plus nécessaires à la non-phrase qu'elle « n'est pas autonome ».

Les manuels d'écriture et de stylistique (Björk *et al*., 1988 ; Fries, 1952, *inter alia*) recommandent d'éviter leur emploi bien qu'elles soient bien représentées dans les textes, comme en témoignent les grammaires (Biber *et al*., 1999 ; Riegel, Pellat, Rioul, 1994, par exemple). Nous analyserons ici les non-phrases thématiques introduites par une marque linguistique particulière : un introducteur thématique qui se manifeste sous la forme d'une préposition simple ou composée.

2.2. Travaux informatiques

Sur le plan informatique, il n'existe pas, à notre connaissance, beaucoup de travaux sur la question. Marcu (2000) donne un seul exemple de non-phrase introduite par un adverbial explicitant une élaboration : *[John likes sweets.][Most of all, John likes ice cream and chocolate.]* [In contrast, *Mary likes fruits.]* [Especially *bananas and strawberries*.].

Certes, les adverbiaux constituent des indices linguistiques importants. Toutefois, dans le cas des introducteurs thématiques, une unité lexicale précédée d'un adverbial peut instaurer plusieurs relations. En fait, en plus de l'adverbial il faut considérer des critères positionnel et ponctuationnel. Grisham (1986) cite trois systèmes (New York University Linguistic String Project, LIFER, le système PLANES de Waltz) qui ont tenté d'adapter à des

non-phrases des techniques utilisées dans des phrases. Ces systèmes reconstruisent les informations manquantes dans les non-phrases en utilisant deux types d'information : (i) une information langagière et du domaine, et (ii) une information contextuelle prenant en compte le discours précédent immédiat. L'auteur souligne les faiblesses de ce type d'approche, fondée sur des propriétés syntaxiques particulières tout en étant étroitement liées à un domaine particulier.

2.3. Une analyse linguistique descriptive plus approfondie

Comme à notre connaissance, les non-phrases n'ont pas fait l'objet d'une étude approfondie, nous avons entrepris de les analyser uniquement d'un point de vue linguistique. Nous exposons ci-dessous les résultats d'une recherche de nature descriptive. Celle-ci liste les critères caractérisant les non-phrases thématiques des non-phrases a-thématiques, c'est-à-dire que dans ce dernier cas, la préposition potentiellement introducteur thématique n'instancie pas de cadre. Cette analyse se fonde sur un corpus écrit composé de 87 extraits de romans (R), de la base de données Frantext, de journaux : *Le Monde Diplomatique, Le Point, Le Figaro* (AJ) et d'articles scientifiques (A). Nous n'évoquerons donc pas les critères typiquement oraux tels l'intonation.

2.3.1. LES PHRASES THÉMATIQUES ET LES NON-PHRASES THÉMATIQUES PARTAGENT-ELLES DES PROPRIÉTÉS ? Avant d'entreprendre l'analyse des non-phrases thématiques nous avons déterminé quelles étaient les caractéristiques des phrases thématiques qui les concernaient : **1)** caractéristiques morphologiques : les non-phrases thématiques sont introduites par des unités lexicales spécifiques. Seules les prépositions listées dans Porhiel (2004) sont concernées ; **2)** caractéristiques syntaxiques : ce sont des phrases sans verbe au complément prototypiquement court et dont la préposition peut se trouver après un adverbial : *- Bien sûr, maître Hsueh, bien sûr, dit vivement Ti./- Mais,* ***à propos de*** *l'historien Shu. Continuez, je vous prie*. (...). (R). Toutefois, à la différence des introducteurs thématiques des phrases thématiques, ceux des non-phrases thématiques ne préfixent pas de proposition(s). De ce fait, le critère syntaxique de la préfixation, fortement discriminant dans les phrases thématiques est neutralisé dans les non-phrases thématiques. Sur la base d'un tel critère, il est impossible d'affirmer si les exemples 1) 2) ou 3) sont des non-phrases thématiques ou a-thématiques ; **3)** caractéristiques ponctuationnelles : l'introducteur suivi de son complément, n'est pas séparé des propositions par une virgule. En revanche, le complément peut être délimité par divers signes de ponctuation : [.], [:], […], [?] ; **4)** caractéristiques sémantiques : les introducteurs des non-phrases thématiques présentent les mêmes caractérisitiques sémantiques que ceux des phrases

thématiques (*Cf.* 1). Les caractéristiques des non-phrases rapidement évoquées ci-dessus soulignent : (i) que les caractéristiques morphologiques ne servent qu'au repérage des non-phrases potentiellement thématiques ; (ii) qu'il n'existe plus de critère opérationnel pour caractériser une non-phrase thématique d'une non-phrase a-thématique. Par conséquent, la prochaine étape devra déterminer les caractéristiques spécifiques des non-phrases par rapport aux phrases, et les caractéristiques spécifiques des non-phrases thématiques par rapport aux non-phrases a-thématiques.

2.3.2. LES CARACTÉRISTIQUES SPÉCIFIQUES AUX NON-PHRASES THÉMATIQUES. Si la phrase et la non-phrase sont définies comme une suite de mots délimitée par une lettre majuscule initiale et par une ponctuation forte finale, rien ne les distingue l'une de l'autre. Il faut par conséquent, trouver d'autres propriétés, sachant que les non-phrases se définissent uniquement à l'aide d'un faisceau de critères, qui n'ont de pertinence réelle que les uns par rapport aux autres : la ponctuation, leur position dans le texte, l'adverbial présent dans la non-phrase et la nature de l'extrait textuel.

Les signes de ponctuation sont une caractéristique propre aux textes écrits et font partie des caractéristiques d'un texte, comme l'ont montré des chercheurs en linguistique (Catach, 1994 ; Védénina, 1989 ; Nunberg, 1990, par exemple) et en traitement automatique des textes (Mourad, 1999 ; Marcu, 1998 et 2000, par exemple). Dans le cas des non-phrases thématiques, la virgule fournit un bon critère de reconnaissance : elle sépare prototypiquement le cadre thématique du reste de la ou des propositions : ***Pour ce qui concerne*** *la France, les organisateurs de la conférence de Stockholm se montrent extrêmement critiques à l'égard du Minitel* (…) (AJ). Dans le cas des non-phrases, les signes de ponctuation [.], […], [:] et [?] initient des cadres thématiques (ouverture à droite) ou a-thématiques (fermeture à gauche et restriction sémantique). De ce fait, on peut simplement dire (i) que 4) et 5) sont des non-phrases : elles n'ont pas de verbe et sont délimités orthographiquement ; (ii) qu'elles sont potentiellement thématiques car elles sont introduites par les prépositions *à propos de* et *pour ce qui est de* :

> 4*) Grâce à la présence, autour de lui [mon père], de sa petite famille humaine et animale, il devait y demeurer vingt-huit ans.* [Nouveau chapitre]
> ***À propos d****'évêque. Je ne sais comment mon père décrocha cette recommandation.* (R)
> 5) *Hélas, mon histoire n'est guère différente.* ***Pour ce qui est de*** *la mort en tout cas.* (R)

Quant aux signes de ponctuation [;] et [!], ils n'initient pas de cadre, au moins dans notre corpus. Dans (6) *en ce qui concerne* introduit une réponse à une question rhétorique. Il s'agit d'un énoncé elliptique :

> *6) Sune Albinsson (...) a monté une société de financement – la Blekinge A.B. – dans le but de créer ou d'encourager les petites entreprises de la région.* (...) *Un pays socialiste, la Suède ? Absolument pas,* ***en ce qui concerne*** *l'industrie !* (...). (AJ)

Pour résumer, le [.], les […], les [:] et le [?] ont une fonction démarcative orthographique : ils indiquent la limite du complément thématique (prototypiquement court et de forme Dét N) dans une non-phrase de la même façon que la virgule dans une phrase thématique. Ainsi : dans une phrase thématique, la limite orthographique du complément thématique est signalée par une [,] ; dans une non-phrase thématique, la limite orthographique du complément thématique est signalée par un [.], […], [:] ou un [?]. Les signes de ponctuation sont nécessaires mais non suffisants pour distinguer les non-phrases thématiques des non-phrases a-thématiques. Néanmoins, ils ont leur importance, combinés à d'autres critères comme celui de position.

Le critère de position doit, pour être opérant, dépasser le cadre de la phrase : dans les non-phrases, les prépositions potentiellement introducteurs thématiques se trouvent en position initiale ou après un adverbial. Thompson et Langacre (1985), Virtanen (1992), Goutsos (1997) *inter alia* ont souligné le rôle, textuel et organisationnel, des groupes adverbiaux de temps et de lieu, préfixés en début de phrase ou en début de paragraphe. Ils mentionnent que les éléments préfixés jouent un rôle organisationnel, ce que ne font pas (ou pas à la même échelle) les groupes prépositionnels en fin de phrase. Si les non-phrases ne suivent pas les règles habituelles de la syntaxe afin de rendre un texte plus vivant, il y a de fortes chances que certaines jouent un rôle indentique à celui des groupes adverbiaux préfixés : certains ouvrent vers la droite et d'autres saturent sémantiquement un élément morpho-syntaxique, c'est-à-dire qu'ils focalisent ou ferment à gauche. La combinaison du critère positionnel en début de paragraphe *vs*. dans le paragraphe, combiné à celui de la ponctuation permet de préciser la nature de la non-phrase dans les exemples : (5) est une non-phrase thématique et (6), une non-phrase a-thématique. En d'autres termes, (5), en début de paragraphe, s'ouvre sur la droite et joue un rôle discursif organisationnel et cohésif. Un telle combinaison (ponctuation+position) fonctionne aussi dans la plupart des cas avec [:] ***À propos de*** *démocratie: Jabotinski se définissait comme un libéral et défendait avec fermeté le système parlementaire* (AJ) vs. *Le bilan en est controversé et son éventuelle révision est maintenant évoquée. Non pas* ***à l'égard de*** *l'Irak: même avec l'application de la résolution 986 (...)* (AJ), et ne fonctionne pas avec […] et [?]. Ainsi dans *Il va falloir que tu loges en ville, chez une dame Polin qu'il recommande.* ***À***

propos de *Ladourd… Voici la barrière qui fut blanche, voici la route goudronnée*… (Frantext), la non-phrase est thématique, bien qu'elle se trouve dans le paragraphe. Néanmoins, les non-phrases ne sont caractérisables qu'avec un faisceau de deux critères au moins. Reste à déterminer quels sont ces critères, et comment ils se combinent aux précédents, ou à un des critères précédents.

Pour arriver à déterminer si une non-phrase est thématique ou a-thématique, on peut prendre en considération la nature de l'adverbial car certains adverbiaux favorisent une lecture thématique des non-phrases : ce sont les marqueurs d'intégration linéaire (MIL) *(d'abord, premièrement, ensuite*, etc.) (Turco, Coltier, 1988), les marqueurs d'interjection (*hé, dis donc*, etc.), et les marqueurs discursifs (*et, justement, mais*, etc.) ; d'autres adverbiaux initient une lecture non thématique : ce sont des marqueurs additifs (*et, ou,* etc.), des marqueurs sélectifs (*notamment*, *tout particulièrement,* etc.), des marqueurs d'exception (*sauf*), des marqueurs de négation (*non pas, ni*, etc.), des marqueurs aspectuels (*toujours*, *jamais*), des marqueurs illustratifs (*par exemple*, etc.), et des réponses courtes (*oui, non, absolument pas*). La prise en compte de ce critère suggère trois remarques : **1)** la nature de l'adverbial (ceux qui favorisent une lecture thématique de la non-phrase) prévaut sur la position de la non-phrase dans le texte comme le montre l'exemple (7) :

> 7) *En réalité, les Australiens de souche sont inquiets pour leur avenir.* ***Sur le plan de*** *la natalité d'abord : dès l'an 2000, leur population va baisser (...).* ***Sur le plan des*** *valeurs ensuite : (...) les Australiens sont devenus blasés* (…). (AJ)

Bien que les prépositions *sur le plan de* se trouvent dans le paragraphe et dans la continuité syntaxique de la phrase introductrice (*En réalité, les Australiens de souche sont inquiets pour leur avenir)*, *sur le plan de* est un introducteur thématique. Comme la non-phrase comporte un MIL dont la fonction est d'organiser le discours, la combinaison introducteur thématique+MIL confère à la non-phrase un rôle de sériation de l'information ; **2)** les adverbiaux sont polycatégoriels et peuvent, comme c'est le cas des MIL être empruntés à d'autres sous-systèmes (Turco, Coltier, 1988 ; Adam, Revaz, 1989 ; Jackiewicz, 2002). Par conséquent, les catégories d'adverbiaux mentionnées initient potentiellement une lecture thématique. Par exemple, *d'abord* peut être un MIL mais aussi un adverbial temporel, *et* peut être un marqueur discursif mais aussi un marqueur additif ; **3)** cet exemple montre aussi que pour analyser les non-phrases thématiques, une phrase adjacente ne suffit pas. Ici, les non-phrases thématiques exemplifient l'argument avancé par la phrase introductrice.

Enfin, nous considérons un dernier critère : extrait dialogal *vs*. extrait non dialogal. D'après le critère de position de la non-phrase dans le texte,

lorsqu'une préposition est potentiellement un introducteur thématique, elle se trouve en début de paragraphe ou de réplique. Dans les dialogues, les prépositions en début de réplique, après un tiret cadratin, introduisent le plus souvent une réponse à une question ou une demande de précision :

8) *Il ajouta un peu plus haut, comme Joseph traversait la rue :*
– pense à la réponse que tu dois me donner.
Joseph se retourna.
– quelle réponse ?
– ***au sujet de*** *la chambre*. (Frantext)

ou bien encore introduisent une phrase laissée en suspens, interrompues par […] :

9) (...) *Vivian les sermonna d'un ton de maîtresse d'école :*
- Vous traînez ici, passant votre temps à inventer des histoires….
- Je ne dirais pas cela, Viv….
Il y eut un froissement de tissu tandis que Trueblood essayait de recroiser les jambes.
- … ***à propos de*** *la famille de Franco. Sa mère n'est pas une matrone à la moustache noire (...)* (R)

Il s'agit alors d'ellipses et de fragments, donc de non-phrases a-thématiques. Toutefois, dans un exemple comme (10), la préposition n'introduit pas un complément dans la continuité syntaxique et sémantique de la réplique précédente :

10) *J'ai encore une ou deux questions, avec votre permission.*
- Si je connais les réponses, vous les aurez.
- ***À propos de*** *votre père. Un grand nombre de ministres de l'Eglise réformée accordent peu d'importance aux fêtes religieuses, (...)* (R)

Dans (10), la phrase qui précède la non-phrase n'est pas une question, alors que c'était le cas dans (8) : *à propos de* réintroduit un des sujets abordés par les deux locuteurs et le développe. Il s'agit d'un introducteur thématique. Ce critère permet d'avancer que : (i) dans un dialogue, l'unité lexicale en début de réplique n'introduit pas un cadre thématique si la réplique précédente se termine par un point d'interrogation (donc s'il s'agit d'une question) (8) ou par trois points (9) ; (ii) dans un dialogue, la préposition en début de réplique introduit un cadre thématique si la phrase précédente se termine par un point. Ce dernier point demande à être confirmé.

À l'issue de l'analyse linguistique des non-phrases dans un corpus écrit, on constate que : (i) les quatre critères utilisés pour caractériser les non-

phrases thématiques correspondent à des emplois particuliers (thématique ou a-thématique) des non-phrases ; (ii) les critères se combinent par deux et les non-phrases sont thématiques ou a-thématiques en fonction de l'absence ou de la présence d'un adverbial.

Les combinaisons caractérisant les non-phrases thématiques se résument alors ainsi :

[P]	dans §	début §	Adverbial
[.]		x	
[.]			x
[:]		x	
[:]			x
[…]	x		
[…]		x	
[…]			x
[?]		x	
[?]			x

Si ce tableau rend compte de la situation dans la non-phrase, il n'est pas suffisant pour décrire les non-phrases dans un dialogue. En effet, dans un dialogue, la réplique précédant la non-phrase thématique ne doit pas se terminer par [?].

3. LES NON-PHRASES THÉMATIQUES DANS LA PLATE-FORME CONTEXTO

3.1. Capitalisation des données

Les données utilisées dans ContextO résultent d'une analyse linguistique approfondie (*cf.* 2.3.). Les propriétés morphologiques, morphosyntaxiques et sémantiques mises en évidence dans Porhiel (2001a/b, 2004) sont réutilisables dans le projet actuel. Ces propriétés ont notamment servi à répartir les introducteurs thématiques, qui déclenchent les règles de répérage, en 27 sous-classes dans la base de données Nous avons également réutilisé les classes d'adverbiaux utiles à la description des non-phrases thématiques.

D'autres données, particulières aux introducteurs thématiques dans des non-phrases, doivent être utilisées dans les règles déclaratives écrites en JAVA. Ces données sont : **a)** syntaxiques : les compléments ne sont pas séparés de la ou des propositions qu'ils indexent par une virgule et sont prototypiquement courts ; **b)** ponctuationnelles : les signes de ponctuation qui constituent la limite droite de la non-phrase thématique sont le [.], les [:], les […], le [?] ; **c)** positionnelles : les non-phrases thématiques se trouvent en

début de paragraphe ou dans le paragraphe ; **d)** lexicales : certains adverbiaux (avant ou après la préposition) favorisent une lecture thématique des non-phrases. Les règles concernant les introducteurs thématiques dans des non-phrases montrent que leur repérage morphologique, combiné à des indices positionnels, textuels, ponctuationnels, lexicaux et syntaxiques inhibent ou favorisent l'extraction des non-phrases qu'ils introduisent. Elles permettent, par exemple, de repérer les non-phrases thématiques en début de paragraphe, après un adverbial. Ces règles sont ensuite associées aux propriétés morpho-syntaxiques et sémantiques des introducteurs thématiques. L'opération est reproduite pour chacune des catégories d'introducteurs thématiques (27 fois). L'écriture de telles règles fondées sur une analyse linguistique a soulevé des problèmes confrontant la réalité linguistique et la réalité informatique. On peut notamment se demander dans quelle mesure il est possible d'implémenter informatiquement des données linguistiques.

3.2. Réalité linguistique et réalité informatique ?

Lors de l'étape qui consiste à implémenter des données linguistiques dans un programme informatique, le chercheur se trouve confronté à un problème de *traduction* : comment rendre fidèlement dans un langage de nature différente (humain *vs*. machine), de syntaxe différente (« liberté d'expression » *vs*. binarité et espaces de recherche) des informations linguistiques. Nous évoquons quatre cas de figure qui confrontent linguistique et informatique : la limite orthographique du complément, la longueur des compléments, les insertions et la polycatégorialité des adverbiaux.

- La limite orthographique du complément : protoypiquement, dans une phrase, la virgule sépare l'introducteur thématique et son complément, de la proposition qu'ils indexent, tandis que dans la non-phrase, la limite du complément est marquée par [.], [:], […], [?]. Une telle propriété permet de repérer des phrases ou des non-phrases introduites par une préposition, potentiellement introducteur thématique. Les règles spécifient si le complément est clôturé par une [,] ou par [.], [:], […], [?]. Toufefois, il peut arriver qu'une virgule se trouve dans une non-phrase, quand le complément introduit par l'introducteur est expansé, ou quand le complément est suivi d'un adverbial : *À propos de Mai 68, encore – et en vrac. Et si la vrai commémoration, la seule, nous venait d'Indonésie (...)* (AJ), ce qui soulève le problème de la longueur des compléments introduits par l'introducteur, et aussi le fait que l'adverbial peut suivre le complément thématique et en être séparé par une virgule.
- La longueur du complément : Dans une non-phrase thématique, les compléments introduits par l'introducteur sont prototypiquement courts et de forme : N, Dét N, Dét N Adj., Dét. Adj. N, voire Dét. Adj. N Adj. Sur le

plan informatique, en termes numériques (espaces de recherche), cela nous ramène à un complément composé de quatre mots (au sens orthographique) et dont la limite orthographique, marquée par [.], [:], […], [?], se trouve en cinquième position. Toutefois, les compléments peuvent aussi, bien que rarement, être expansés par une coordination, un subordonnant ***À propos de*** *la manifestation palestinienne qui s'est déroulée place de la République, à Paris : Il est intolérable que* (...) (AJ), ou par d'autres informations notamment dans les rubriques « courrier des lecteurs » : ***À propos de*** *l'édit de Nantes (n° 1325). On a coutume de dire que* (...) (AJ). Dans le deuxième cas, les informations fournies sont du même type (renvoi au numéro du journal, au titre de l'article, à la date de parution, etc.) et il est possible de les coder sous forme de patrons morpho-syntaxiques et sémantico-lexicaux, clôturés par [.], [:], […], [?]. Dans le premier cas, l'entreprise est plus aléatoire. On peut néanmoins indiquer sous forme de patrons morpho-syntaxiques la constitution de l'expansion : subordonnant, déterminant, nom, plus la clôture [.], [:], […], [?]. Dans les règles que nous avons écrites nous avons envisagé deux cas de figure : un dans lequel le complément n'est pas expansé, et un autre dans lequel le complément est expansé.

- Les insertions : Les introducteurs thématiques acceptent des insertions, c'est-à-dire des éléments facultatifs. Notre corpus ne comporte pas d'exemple concernant les introducteurs thématiques dans les non-phrases, ce qui ne signifie pas qu'une telle configuration soit impossible. Elle a donc été prise en compte et, pour ce faire, nous avons utilisé les résultats de Porhiel (2001a) : la longueur des insertions autorisées dans les prépositions composées se limitent à 3 mots.
- La plupart des adverbiaux instaurent par nature différents types de relations. Selon la nature de l'adverbial, une non-phrase aura une lecture thématique ou a-thématique. La désambiguïsation des adverbiaux, dépassant le cadre de ce projet, nous n'avons pas cherché à « résoudre les erreurs de repérage ».

Nous venons d'évoquer quatre cas de figure de données linguistiques qui s'écartent dans l'absolu des propriétés prototypiques des non-phrases thématiques, expliquées au point 2.3. Sur le plan informatique, les propriétés prototypiques présentent l'avantage d'être bornées (ici par exemple, la limite orthographique, les insertions) et de pouvoir être traduites en données chiffrées si besoin (ici par exemple, la longueur des compléments). Par ailleurs, les données prototypiques satisfont à la fois le linguiste et l'informaticien, ce qui n'est pas toujours le cas avec les configurations non prototypiques. Dans ce dernier cas, la réalité linguistique et la réalité informatique ne trouvent pas toujours facilement un terrain d'entente : quel seuil d'imprécision le linguiste est-il prêt à tolérer ? Ce qui a pour corollaire informatique le degré de robustesse des règles de repérage. Il faut aussi répondre à certaines questions : jusqu'à quel point prendre en compte les

propriétés non prototypiques ? quel est le coût de cette prise en compte ? La solution envisagée est-elle trop *ad hoc* car elle s'aligne sur le corpus analysé ? À ce sujet, il faut faire preuve d'honnêteté intellectuelle et reconnaître que toute solution envisagée est liée à un corpus et s'insère dans la logique d'un projet : une partie de projet ne remet pas normalement en question l'ensemble du projet.

La réalité linguistique et la réalité informatique suggèrent que les deux disciplines ont une logique de travail. Il nous semble que la rencontre entre la linguistique et l'informatique rappelle le travail du traducteur. En passant des données de l'analyse linguistique à l'implémentation informatique, on déclenche des processus cognitifs qui ont une réalité linguistique et qui ont besoin d'être explicités, (décodés, en somme) traduits, voire réinterprétés/transformés, si on désire obtenir des résultats de repérage satisfaisants. Par exemple, la marque initiale du linguiste est réinterprétée/transformée en token en position 1. Une telle réinterprétation/transformation, contraint le linguiste (i) à présenter des données de façon systématique en constituant des listes (critères de reconnaissance extensifs) ou en spécifiant des patrons morpho-syntaxiques (critères de reconnaissance intensifs) et (ii), surtout à mieux cerner des catégories, par exemple celle des adverbiaux qui posent également problème en linguistique. Le détail informatique peut être sans importance sur le plan linguistique tout comme le détail linguistique peut être non nécessaire sur le plan informatique. Les problèmes rencontrés permettent alors de questionner, sans toutefois remettre en question, la « validité » de certaines données linguistique dans l'extraction automatique.

3.3. Extraction des non-phrases thématiques : résultats et discussion

L'écriture des règles permettant de repérer les non-phrases thématiques s'est faite en trois étapes, selon un processus dialogal et incrémentiel.

- Première étape (implémentation des données linguistiques) : les règles ont été écrites à partir des données de l'analyse linguistique. Elles ont ensuite été lancées sur un corpus et les résultats obtenus ont donné lieu à une deuxième phase dans l'implémentation ;
- Deuxième étape (dialogue réalité linguistique-réalité informatique) : les règles de la première étape ont été modifiées en tenant compte des paramètres informatiques. Il a notamment fallu interpréter, traduire les données linguistiques en langage informatique. C'est lors de cette phase qu'ont été définies la longueur des insertions dans les prépositions composées et la longueur des espaces de recherche des compléments, de façon à optimiser les résultats. Cette étape peut s'avérer délicate. Ainsi, un linguiste ne s'attarde pas sur la longueur d'un complément. Comment

d'ailleurs trouver la longueur « idéale » d'un complément ? Quelles sont les implications de données fixes ? Il faut donc accepter qu'une règle « présentera des lacunes » car elle ne considère pas des configurations peut récurrentes ;

- Troisième étape (dialogue règles-corpus) : la troisième étape consiste : (i) à évaluer (évaluation qui peut par la suite être revue) le coût, la pertinence de certaines modifications dans les règles, (ii) à ne pas chercher à implémenter des configurations peu courantes.

Ces trois étapes ont permis d'écrire des règles, aux configurations syntaxiques différentes, dans la plate-forme ContextO. Les septs catégories de règles ci-après permettent de repérer les non-phrases thématiques : en début de paragraphe (non expansé et expansé) ; après un tiret (non expansé et expansé) ; après un tiret et un adverbial (non expansé et expansé) ; après un adverbial (non expansé et expansé) ; après un adverbial suivi d'une virgule (non expansé et expansé) ; suivi d'un adverbial (non expansé et expansé) ; suivi de trois points (non expansé et expansé).

Le corpus sur lequel se fonde notre analyse est composé de 85 extraits (les non-phrases ne sont donc pas isolées) de journaux et de romans. Nous y avons relevé manuellement 88 occurrences d'unités lexicales qui ont la potentialité d'être des introducteurs thématiques. En fait, 33 (37,5 %) de ces unités lexicales introduisent des non-phrases thématiques et 55 (62,5 %) des non-phrases a-thématiques. Les résultats montrent que les segments thématiques du corpus ont été dans l'ensemble correctement repérés : 1) sur 33 unités lexicales qui introduisent des non-phrases thématiques, 4 n'ont pas été reconnues : (i) les configurations sont peu récurrentes ou (ii) les règles n'ont pas encore été écrites ; 2) sur 55 unités lexicales qui n'introduisent pas des non-phrases thématiques, 9 ont été extraites. Toutefois, ce ne sont pas les règles concernant les non-phrases thématiques qui ont été déclenchées, mais les règles concernant les non-phrases a-thématiques : (i) dans 6 cas, les compléments sont expansés par une coordination, une subordonnée ou une apposition. Ainsi, le subordonnant étant précédé d'une virgule, le programme informatique indique que la non-phrase est en fait une « phrase thématique » : *De nombreux cas exemplaires de cette interdépendance planétaire ont été mis en évidence lors du congrès* (...). *En particulier* ***à propos de*** *la destruction des forêts tropicales, qui pourtant représentent un précieux réservoir biologique pour la santé (...)* (AJ). Ce point suggère qu'il faudra sans doute prendre aussi le critère positionnel pour repérer les phrases thématiques, ce qui n'est pas actuellement le cas. En tout état de cause, les règles concernant les non-phrases fonctionnent. (ii) Dans 2 cas, le programme reconnaît des non-phrases thématiques alors que ce sont des phrases thématiques. Dans *Vous cherchez un point de vente près de chez vous, ou bien vous souhaitez connaître l'actualité de votre région ? Rien de*

plus simple avec notre cahier qui regroupe plus de 3 500 revendeurs de proximités. ***Quant à*** *nos fameuses promotions… vous en avez près de 500 à étudier ce mois-ci.* (…) (AJ), les […] moins neutres d'un point de vue interprétatif sont utilisés en lieu et place de la virgule ; le premier mot de la proposition commence par une minuscule et non par une majuscule (sinon on a affaire à une non-phrase thématique. (iii) Dans un cas, une non-phrase a-thématique a été reconnue.

Ces règles ont aussi été testées sur un corpus (corpus b), composé de 7 extraits, comportant 12 unités lexicales potentiellement introducteurs thématiques (6 introducteurs (50 %) et 6 qui ne sont pas des introducteurs (50 %)) dans des non-phrases, et n'ayant pas servi à l'analyse précédente. Les résultats sont similaires à ceux obtenus précédemment. En récapitulant nous obtenons la répartition chiffrée suivante :

	Corpus a		Corpus b
Nombre d'introducteurs correctement repérés / nombre d'introducteur présents dans le texte (rappel)	84,8 %		50 %
Nombre d'introducteurs correctement repérés/ nombre d'introducteurs repérés (précision)	93,3 %		87,5 %

Ces données encourageantes devront être vérifiées sur un corpus plus large que celui utilisé dans cette étude.

Dans ce travail nous avons défini les critères qui permettent de caractériser les non-phrases thématiques. Aucun des critères considérés n'est isolément déterminant mais leur interaction fournit des critères pertinents. Ces données ont permis d'écrire des règles de repérage pointant sur la structure thématique d'un extrait, d'un passage de texte. Les résultats obtenus sont prometteurs et pourront être utilisés pour repérer les structures discursives que ces introducteurs thématiques engendrent comme dans (11) :

> 11) *(…) Nous formulons ci-dessous quelques uns des présupposés les plus évidents de la RST, concernant le langage, les observateurs, les scripteurs et l'interprétation des régularités dans les séries d'analyses.*
> ***Concernant*** *le langage : la RST présuppose qu'un texte monologal (…)*
> ***Concernant*** *les observateurs : nous présupposons que (…)*
> ***Concernant*** *les scripteurs : nous présupposons que (…)*
> ***Concernant*** *les régularités dans les séries d'analyses : les notions de texte (…)*

Une telle démarche s'appuiera sur les données des analyses de Jackiewicz, Minel (2003) obtenues lors du repérage des structures discursives des cadres organisationnels.

REFERENCES

Adam, J.-M., Revaz, F. 1989. "Aspects de la structuration du texte descriptif : les marqueurs d'énumération et de reformulation", *Langue Française*, n°81, p. 59-98.

Ben Hazez, S., Minel, J-L. 2000. "Designing Tasks of Identification of Complex Patterns Used for Text Filtering", *RIAO'2000*, p. 1558-1567.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. 1999. *Longman grammar of spoken and written English*. London : Longman.

Björk, L., Knight, M., Wikborg, E. 1992. *The Writing Process – Composition Writing for University Students*. 2ième édition, Lund : Studentlitteratur.

Catach, N. 1994. *La ponctuation*, Paris : PUF (« Que Sais-je ? », n° 2818).

Charlolles, M. 1997. *L'encadrement du discours : univers, champs, domaines et espaces*, Cahier de Recherche Linguistique, LANDISCO, URA-CNRS 1035 Université Nancy 2, n° 6, p. 1-73, [Accessible sur le site http://www.ltm.ens.fr/siteACFT/].

Dea, W., Belkin, N. J. 1978. "Beyond the sentence : clause relations and textual analysis", K. P. Jones, V. Horsnell (éds.), *Informatics*, n°3, London, Aslib, p. 67-83.

Descles, J-P., Cartier, E., Jackiewicz, A., Minel, J.-L. 1997. "Textual Processing and Contextual Exploration Method", *CONTEXT 97*, Universidade Federal do Rio de Janeiro, Brésil, p. 189-197.

Ferret, O., Grau, B., Minel, J.-L., Porhiel, S. 2001. "Repérage de structures thématiques dans des textes", *TALN 2001*, Tours 2-5 juillet 2001, p. 163-172.

Fries, C. C. 1952. *The Structure of English - An Introduction to the Construction of English Sentences*. New York : Harcourt.

Goutsos, D. 1997. *Topic : sequential relations and strategies in expository text*, *Advances in discourse proce*sses vol. 9.

Grisham, R. 1986. *Computational Linguistics – An introduction*. Cambridge : University Press.

Hollerbach, W. 1994. *The Syntax of Contemporary French – A Pedagogical Handbook and Reference Gr*ammar. New York : University Press of America.

Jackiewicz, A. 2002. "Repérage et delimitation des cadres organisationnels pour la segmentation automatique des texts", *CIFT'02*, Hammamet, Tunisie, p. 95-107.

Jackiewicz, A., Minel, J.-L. 2003. "L'identification des structures discursives engendrées par les cadres organisationnels", *TALN 2003*, Batz-sur-Mer, 11-14 juin 2003.

Knott, A., Sanders, T. 1998. "The Classification of Coherence Relations and their Linguistic Markers : An Exploration of Two Languages", *Journal of Pragmatics*, n°30, p. 135-175.

Mann, W., Thompson, S. 1988. "Rhetorical Structure Theory : Toward a Functional Theory of Text Organization", *Text*, n°8, p. 248-281.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge : The MIT Press.

Marcu, D. 1998. "A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts", COSTERMANS, FAYOL (éds.), *Workshop on Discourse Relations and Discourse Markers*, COLING/ACL'98 Montréal Canada, p. 1-7.

Minel, J.-L. 2003. *Filtrage sémantique. Du résumé à la fouille de textes*. Paris : Hermès.

Minel, J.-L., Cartier, E., Crispino, G., Descles, J.-P., Ben Hazez, S., Jackiewicz, A. 2001. "Résumé automatique par filtrage sémantique d'informations dans des textes", *Technique et Science Informatiques*, n°3, Paris, n°3, p. 369-395.

Mourad, G. 1999. "La segmentation des textes par l'étude de la ponctuation", *CIDE'99*, Damas, Syrie.

Nunberg, G. 1990. *The Linguistics of Punctuation*, Center for the Study of Language and Information, Standford, Calif.

Porhiel, S. 2001a. "Organizing Linguistic Data : Thematic introducers as an Example", Coyote Papers, n°12, p. 42-61.

Porhiel, S. 2001b. "Linguistic expressions as a tool to extract thematic information", P. Rayson et al. (éds.), *Corpus Linguistic 2001*, Lancaster University, Lancaster du 31 mars au 2 avril 2001, Technical Papers Volume 13 – Special issue, p. 477-482.

Porhiel, S. 2004. "Les introducteurs thématiques", *Cahiers de Lexicologie*, n° 85, 2, p. 9-45.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. New York : Longman.

Riegel, M., Pellat, J.-C., Rioul, R. 1994. *Grammaire méthodique du français*. Paris : PUF.

Thompson, S., Langacker, R. 1985. "Adverbial Clauses", SHOPEN (éd.), *Language Typology and Syntactic Decription - Complexe Construction*, vol. 2. Cambridge : Cambridge University Press. p. 170-234.

Turco, G., Coltier, D. "Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire", *Pratiques*, n°57, p. 57-79.

Vedenina, L. 1989. *Pertinence de la présentation typographique*. Paris : Peeters/Selaf.

Virtanen, T. 1992. *Discourse Functions of Adverbial Placement in English*. Åbo : Åbo Akademi University Press.

Winter, E. 1978. "A Look at the Role of Certain Words in Information Structure", K. P. Jones, V. Horsnell (éds.), *Informatics*, n°3, Cambridge, London, Aslib, p. 85-97.

NOTES

(1) Je remercie Jean-Luc Minel pour sa relecture de l'article et ses suggestions.

APPORT DE L'ANALYSE LINGUISTIQUE POUR L'EXTRACTION TERMINOLOGIQUE EN CORPUS : APPLICATION AU DOMAINE DE LA GÉNOMIQUE.

Fabienne Ville-Ometz, Alain Zasadzinski, Dominique Besagni
INIST-CNRS

1. INTRODUCTION

L'extraction de la connaissance, exprimée principalement sous forme langagière dans les documents, repose sur des techniques issues du TALN (traitement automatique du langage naturel) et de la terminologie computationnelle. La recherche en corpus des termes porteurs de l'information pertinente est réalisée par notre plate-forme d'ingénierie linguistique ILC − Infométrie, Langage, Connaissances (Royauté, 1999) − qui reconnaît et extrait les termes d'un texte à partir d'une ressource terminologique de référence (section 2). Nous proposons d'analyser la procédure d'indexation automatisée réalisée à l'aide d'ILC et de nous focaliser sur la reconnaissance des variantes terminologiques gérée par l'analyseur FASTR (Jacquemin, 1997) intégré à notre plate-forme. Cet outil effectue un traitement local du groupe nominal à partir de patrons morpho-syntaxiques et utilise un ensemble de règles spécifiques, appelées « métarègles », qui permettent d'identifier les différentes variations linguistiques que peut subir un terme en corpus. Or, les métarègles génèrent des erreurs récurrentes dans la reconnaissance des variantes syntaxiques qui sont liées à des problèmes de dépendances non préservées. Ce manque de précision dans la reconnaissance de la variation nécessite alors que l'ensemble des transformations produites passe par une validation humaine. Partant de ce constat, nous souhaitons intervenir sur ces règles afin d'en renforcer le pouvoir filtrant, mais également afin de diminuer le temps consacré à la validation, et assister le spécialiste dans cette tâche en produisant une indexation la plus fiable possible d'un point de vue linguistique. Pour y parvenir, nous nous appuyons sur une analyse

linguistique des données issues d'une indexation automatique réalisée dans le cadre d'un processus de fouille de données textuelles, appliqué à la génomique du cancer de la thyroïde[1] (section 3). L'étude de l'ensemble des variations produites lors de cette procédure (section 4) a fait émerger deux principaux phénomènes linguistiques à l'origine de mauvaises variations : d'une part, des erreurs de délimitation de la séquence textuelle en corpus et, d'autre part, des rapports de dépendance modifiés par insertion d'une unité linguistique, lexicale et/ou grammaticale pendant la transformation (section 4.2.). Cette étude a également permis de mettre en évidence le comportement de l'expert face à la validation de l'indexation qui traduit de la volonté de produire de l'information implicite (section 4.1.).

À partir de cette analyse, nous proposons un ensemble de critères morphologiques (flexionnels et dérivationnels) et syntaxiques visant à maintenir les relations de dépendance initiales lors de la transformation et à augmenter ainsi la proportion des variations syntaxiques linguistiquement valides (section 5) extraites par le système.

2. TRAITEMENTS LINGUISTIQUES POUR L'EXTRACTION TERMINOLOGIQUE EN CORPUS

2.1. Reconnaissance *versus* acquisition terminologique

Il est couramment admis que les termes sont les représentants linguistiques des concepts, des « étiquettes de concept » (Bourigault et Jacquemin, 2000) et forment des indicateurs privilégiés de la connaissance portée par les documents. Partant de ce postulat, l'extraction de termes dans les textes constitue, depuis le début des années 90, une problématique de recherche et de développement très riche, tant dans le domaine du TALN que dans celui de la terminologie computationnelle. Du point de vue des applications, la recherche de termes dans les textes représente également un enjeu important par rapport à des domaines aussi variés que la recherche d'information, la création de lexiques, dictionnaires ou ressources terminologiques, la traduction automatique, l'aide à la traduction, l'acquisition de connaissances à travers les ontologies, la gestion documentaire, etc. De nombreux outils d'extraction terminologique ont ainsi été développés, dont la majorité repose sur une approche par acquisition de candidats-termes tel que Termino (David et Plante, 1990), application pionnière dans ce domaine, suivi rapidement d'autres systèmes tels que Acabit (Daille, 1994), Lexter (Bourigault, 1994), Ana (Enghenhard, 1992), Xtract (Smadja, 1993). Un large tour d'horizon de ces différents travaux est proposé par Jacquemin (1997) et Bourigault *et al.* (2000, 2001). L'approche par acquisition de termes à partir de corpus concerne principalement la construction de ressources terminologiques représentant la connaissance d'un domaine et dédiées à une application

spécifique (Rastier, 1995). L'approche par reconnaissance terminologique permet de retrouver et d'extraire en corpus de la connaissance déjà acquise et stockée sous forme de base terminologique. Extraire l'information dans les textes par reconnaissance terminologique présente l'avantage d'obtenir une indexation de meilleure qualité par rapport à une indexation par extraction libre. Elle se révèle moins bruitée, plus pertinente et, surtout, plus homogène. La qualité de l'indexation se répercute au niveau des traitements postérieurs dédiés à une analyse des données basée sur des méthodes symboliques et/ou statistiques (Polanco *et al.*, 2000). Toutefois, une des difficultés auxquelles nous sommes confrontés dans ce type d'approche est l'absence de figement d'un terme motivé linguistiquement. Bien que, dans le langage scientifique et technique, le terme se caractériserait par une tendance au figement et à la lexicalisation, il reste néanmoins sujet à certaines variations. Les termes peuvent se réaliser en corpus sous différentes variantes, en premier lieu, parce qu'ils subissent les contraintes de la langue tout comme les groupes nominaux ordinaires, mais également en raison d'un manque de consensus entre les experts. La variation serait alors révélatrice de l'évolution des concepts scientifiques et techniques (Ibekwe-SanJuan 1998 ; Daille *& al*. 1996).

Le principe de la variation conserve la sémantique du terme de départ et renvoie ainsi au même concept. En d'autres termes, le sens porté par la forme variante doit nécessairement impliquer le sens véhiculé par le terme sous sa forme canonique. C'est sur cette approche que s'appuie l'analyseur syntaxique FASTR, développé par Jacquemin (1997), qui reconnaît et extrait les termes ainsi que leurs variantes en corpus. Cet outil permet d'identifier quatre types de variation : les variations flexionnelles, syntaxiques, morpho-dérivationnelles et sémantiques. Gérer la variation terminologique présente un réel avantage, non seulement dans une approche d'analyse de l'information et de fouille textuelle, mais également en informatique documentaire pour la classification des documents (Ibekwe-SanJuan, 1998), l'expansion des requêtes en recherche d'information (Rinaldi *et al*., 2002), ou encore dans le domaine de la construction et de la structuration automatique de terminologie (Morin *et al.* 1999 ; Daille 1994).

2.2. Architecture de la plate-forme ILC

L'approche mise en œuvre dans notre plate-forme d'ingénierie linguistique permet une indexation automatique contrôlée. ILC retrouve dans les textes les termes déjà connus et répertoriés dans un référentiel terminologique. Cette approche s'appuie principalement sur des traitements linguistiques, de type morphologique et syntaxique, et utilise des ressources extérieures (lexicales et terminologiques). Elle permet à un opérateur humain de traiter l'information contenue dans un ensemble de textes sans avoir à lire les

documents de manière séquentielle. La plate-forme repose sur le postulat maintenant bien établi que l'information est véhiculée de manière privilégiée par des groupes nominaux plus ou moins complexes.

La plate-forme ILC constitue un environnement ouvert pour le traitement du langage naturel. Elle accueille et intègre dans des shells UNIX un ensemble d'outils TALN et de ressources linguistiques pour la reconnaissance des termes et de leurs variantes en anglais à partir du texte intégral. Elle s'appuie sur une démarche onomasiologique, en partant du concept pour rechercher toutes ses expressions linguistiques dans la langue de spécialité.

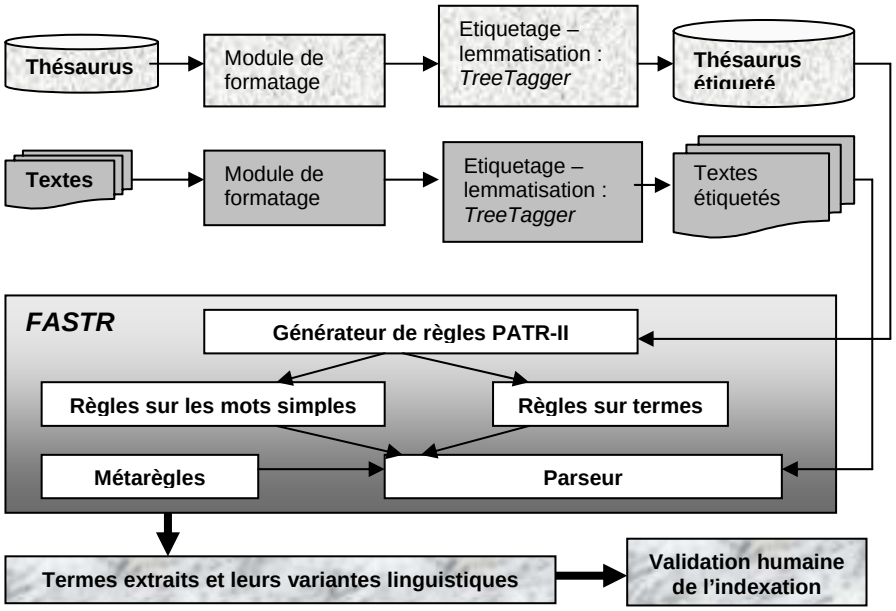


Figure 1 : Architecture de la plate-forme ILC

Ainsi que le montre la figure 1, l'identification des termes et de leurs variantes repose sur deux grandes phases de traitement.

La première phase concerne le traitement des ressources terminologiques. Les termes (termes simples et complexes) sont étiquetés (information grammaticale) puis lemmatisés avec le TreeTagger (Schmid, 1994). Sur ces informations, le programme FASTR, qui utilise le formalisme PATR-II et repose sur une grammaire d'unification, génère un ensemble de règles sur les mots simples composant les termes et sur les termes eux-mêmes. Prenons

l'exemple du pluriterme « residual tumor » enregistré dans le thésaurus de l'UMLS utilisé dans cette expérimentation :

word 'residual' : <cat> = A <root> = ('residue', N)
word 'tumor' : <cat> = N <root> = ('tumor', N)

Les informations concernant les liens morpho-dérivationnels des mots avec leur famille morphologique sont extraites de la base CELEX[2] (base de données lexicale conçue par le « Centre of Lexical Information, Max Plank Instutitute for Psycholinguistics, Nijmegen, Dutch »).

Rule N1 → A2 N3
<N1 lex>='N3'
<N1 lab>='008590'
<A2 lem>='residual'
<N3 lem>='tumor'

La deuxième phase de traitements linguistiques porte sur le corpus qui est également étiqueté et lemmatisé avec TreeTagger puis transformé en PATR-II.

Ces deux types de données (termes et corpus) servent ensuite de données d'entrée à FASTR, qui extrait les termes et leurs variantes textuelles par une analyse locale du groupe nominal à partir de patrons syntaxiques. L'analyseur intègre une métagrammaire en anglais qui permet d'identifier l'ensemble des variations linguistiques autorisées par le système. La plate-forme ILC exploite trois types de variations :

1. la variation flexionnelle : genre et nombre
2. la variation syntaxique

d'insertion : *neural tissue* ← *neural crest derived tissues*
de permutation : *metabolism studies* ← *studies of iodine metabolism*
de coordination : *residual tumor* ← *residual, recurrent or metastatic tumors*

3. la variation morpho-dérivationnelle : *hormone production* ←*produce some others hormones*.

Les métarègles définies dans la métagrammaire autorisent des transformations linguistiques uniquement sur des pluritermes composés de deux ou trois unités (« tumor cells », « the cell », « thyroid function test », « cell of bone »). Ainsi, la métarègle de coordination, formulée à partir de l'expression régulière suivante : X2 N3 → X2 PUNC4 < {A|N|Np|V} PUNC? > C5 < {A|N|Np|V} > N3 va permettre de retrouver en corpus la variante syntaxique *residual, recurrent or metastatic tumors* à partir du terme « residual tumor » enregistré dans l'UMLS.

La première expression indique que le terme de la ressource est composé de deux unités lexicales X2 et N3, appartenant, respectivement, à n'importe quelle partie du discours et à la classe des substantifs. La seconde expression

exprime toutes les possibilités transformationnelles du terme, avec insertion d'une coordination, de ponctuations (la seconde étant optionnelle) et d'unités adjectivales, nominales, verbales ou correspondant à un nom propre.

Les métarègles sont très permissives afin de privilégier le rappel sur la précision. Elles génèrent ainsi des erreurs de variation de manière récurrente. La séquence textuelle ramenée par l'analyseur n'implique pas toujours, d'un point de vue strictement sémantique, le concept auquel renvoie le terme enregistré sous sa forme canonique dans le référentiel (blood vessel ↜ blood flow through selected vessels).

3. CORPUS D'ÉTUDE : LES DONNÉES INITIALES

L'indexation a porté sur 6 256 données bibliographiques (champs textuels des titres et des résumés) en langue anglaise, issues de la base de données bibliographiques biomédicales Medline (Royauté *et al.*, 2004).

La ressource terminologique utilisée pour l'indexation contrôlée se compose de 360 281 termes provenant de l'UMLS[3] (256 290 préférentiels ou concepts et 103 991 synonymes) et susceptibles de subir les variations linguistiques du langage naturel. De cette indexation n'ont été retenus que les termes récupérés à partir d'une variation syntaxique, soit 10 007 séquences textuelles réparties comme de façon suivante :

Variations	Acceptées	Refusées	Total
Insertion	3050 (60,4%)	2001	5051(50.5%)
Permutation	2758 (82,2%)	596	3354 (33,5%)
Coordination	1268 (79,1%)	334	1602 (16%)
Total	7076 (70,7%)	2931	10007

Tableau 1 : Répartition des séquences textuelles selon le type de variation

Dans le cadre de ce travail, nous avons exclu de notre problématique la variation morpho-dérivationnelle. En effet, les erreurs repérées lors de ce type de transformations ne sont pas uniquement imputables à une mauvaise syntaxe des métarègles, mais proviennent partiellement de la base lexicale CELEX à partir de laquelle l'analyseur extrait les informations nécessaires. Cette base fonctionnant sur la notion de famille morphologique au sens large, le passage d'un dérivé à un autre altère souvent le sens de départ (Reproductive system ↜ production in a system).

Chaque type de variations est régi par un ensemble de métarègles. Ainsi, les séquences textuelles obtenues dans cette procédure d'indexation ont été ramenées par 22 métarègles selon la répartition suivante :

Type de variation	Syntaxe de la métarègle	Acceptées	Refusées	Total
Insertion				
XX,16,Ins	X2 N3 → X2 < {A|N|Np|V} 0-3 > N3	2268	1582	3850
XX,19,Ins	X2 N3 → X2 N PREP ART? A? > N3	349	331	680
XX,22,Ins	X2 N3 → X2 < PUNC/')' {A|N|Np|V} ? > N3	122	18	140
XX,25,Ins	X2 N3 → X2 < PUNC/'(' C ? {A|N|Np|V} 1-2 PUNC/')' > N3	97	9	106
XX,28,Ins	X2 N3 → X2 < PUNC/',' {A|N|Np|V} > N3	54	15	69
XXX,28,Ins	X2 A3 N4→ X2 < {A|N|Np|V} 0-3 > A3 N4	5	1	6
XXX,31,Ins	X2 A3 N4→ X2 A3 < {A|N|Np|V} 0-3 > N4	20	2	22
XXX,7,Ins	N2 PREP3 N4 → N2 PREP3< ART ? {A| N|Np|V} 0-3 > N4	135	43	178
Permutation				
XX,31,Perm	X2 N3 → N3 V4 X2 avec <V4 lem> = 'be'	41	22	63
XX,34,Perm	X2 N3 → N3 PUNC4 X2 avec <PUNC4 lem> = '('	116	30	146
XX,37,Perm	N2 N3 → N3 < V ?> PREP4 <ART ? {A|N|Np|V} 0-3 <{N|Np} C ART >? > N2	2445	522	2967
XXX,10,Perm	N2 PREP3 N4 → N4 < {A|N|Np|V} 0-3 N2	156	22	178
Coordination				
XX,1,Coor	X2 N3 → X2 < {A|N|Np|V} 1-3 PUNC/',' ?> C4 < {A|N|Np|V} ? > N3	405	164	569
XX,10,Coor	X2 N3 → X2 PUNC4 < {A|N|Np|V} PUNC5 > < {A|N|Np|V} PUNC? > C6 < {A|N|Np|V} > N3	16	3	19
XX,13,Coor	X2 N3 → X2 < {A|N|Np|V} 0-3 {N|Np} PUNC/',' > C4 N3	0	1	1
XX,4,Coor	X2 N3 → X2 C4 < {A|N|Np|V} 0-3 > N3	736	148	884
XX,7,Coor	X2 N3 → X2 PUNC4 < {A|N|Np|V} PUNC? > C5 < {A|N|Np|V} > N3	76	16	92
XXX,1,Coor	N2 PREP3 N4 → N2 PREP3 < ART? {A|N|Np|V} 0-3 >N6 C5 < PREP? ART? {A|N|Np|V} 0-3 > N4	15	1	16
XXX,19,Coor	X2 A3 N4 → X2 C4 < {A|N|Np|V} 0-3 > A3 N4	6	0	6
XXX,22,Coor	X2 A3 N4 → X2 < {A|N|Np|V} 0-2 > A3 N6 C5 N4	1	0	1
XXX,25,Coor	X2 A3 N4 → X2 A3 C4 < {A|N|Np|V} 0-3 > N4	6	1	7
XXX,4,Coor	N2 PREP3 N4 → N2 C4 < PREP? ART ? { A|N|Np|V} 0-3 > N6 PREP3 N4	7	0	7

Tableau 2 : Type de *métarègles à l'origine des variations dans le processus d'indexation*

Nous remarquons que les métarègles ne présentent pas une productivité identique mais qu'il existe des écarts importants entre elles. Premier constat général, les séquences textuelles (ST) ramenées par des métarègles intervenant sur des pluritermes composés de trois mots sont quantitativement moins importantes que celles issues de bi-termes, alors que le nombre même des métarègles gérant ces deux types de termes est très proche. Cette différence peut s'expliquer de deux manières. D'une part, comme le constate Pozzi (2002), les tri-termes sont généralement moins représentés dans une

terminologie. D'autre part, les possibilités de transformations linguistiques sur ces tri-termes semblent plus réduites, particulièrement lors d'une coordination (Compression of esophagus ← compression of trachea and esophagus ; Electron microscopic study ← electron microscopic and immunohistochemical studies). Les transformations les plus fréquentes correspondent à des insertions (biopsy of thyroid ← biopsy of the human thyroid) et des permutations avec le passage d'une structure à complément du nom introduit par une préposition à une structure préposée (Lesion of brain ← brain lesions). Enfin, en comparant ces résultats à ceux résultant d'une indexation menée dans le domaine de la biologie moléculaire, il est apparu que certaines variations présentent une productivité élevée quelque que soit le domaine d'étude (les métarègles XX, 37, Perm et XX, 16, Ins), alors que d'autres se montrent beaucoup plus sensibles au domaine traité (les coordinations XX, 1, Coor et XX, 4, Coor). Les premières semblent refléter des transformations linguistiques du langage courant, alors que les secondes révèleraient une dépendance plus grande par rapport au langage du domaine concerné.

Pour l'analyse des variations produites par FASTR, nous nous sommes appuyés sur deux corpus d'analyse distincts. Le premier corpus rassemble les données issues de l'indexation : la séquence textuelle, le type de métarègle qui a engendré la variation, le terme ramené par la ST et son synonyme lorsque la variation porte sur lui et qu'il joue le rôle d'interface entre la ST et le terme (*cf*. tableau 2). Le second corpus renvoie aux notices bibliographiques (titres et résumés) et nous a été utile afin de vérifier en contexte la structure linguistique et la sémantique des ST.

Terme	**Synonyme**	**Séquence textuelle**	**Métarègle**
Polyomavirus	Polyoma virus	polyoma murine leukemia virus	XX,16,Ins
Free graft		free pancreas graft	XX,16,Ins
Liver Neoplasm	TUMOR LIVER	tumor development in the liver	XX,19,Ins
Inbred NOD Mice	NOD mouse	nod) mouse	XX,22,Ins
Polyclonal antibody		polyclonal (rabbit) antibody	XX,25,Ins

Tableau 3 : Extrait du corpus de résultats de l'indexation

Le corpus comprenant les résultats de l'indexation a fait l'objet d'une double annotation. La première, de nature linguistique, visait à mettre en valeur des faits récurrents dans le rejet d'une variation. Ceci a permis de faire émerger deux principaux phénomènes linguistiques à l'origine d'une mauvaise transformation, que nous commenterons plus en détail dans la

section suivante : (i) un problème de délimitation de la ST en corpus et (ii) des rapports de dépendance initiaux entre la tête et son expansion, modifiés lors de la transformation. Dans les deux cas, la séquence textuelle et le terme ne renvoient plus au même concept.

Lorsque le rejet de la ST ne s'appuyait pas sur des critères linguistiques, l'expert est intervenu pour effectuer une seconde annotation afin de justifier son choix.

4. VARIATION TERMINOLOGIQUE : DISCUSSION SUR LA DÉMARCHE DE L'EXPERT ET ANALYSE LINGUISTIQUE

4.1. Validation humaine : la démarche de l'expert

Dans la procédure de validation de l'indexation, l'expert n'accède qu'au premier jeu de données et ne peut consulter le résumé. Cette démarche se justifie par le fait qu'il s'agit bien d'un processus d'indexation automatique et non d'une indexation manuelle. Lorsque l'on gère un nombre aussi important de documents, il serait beaucoup trop coûteux en temps de se reporter au texte pour valider chaque terme extrait par le système.

L'étude du corpus doublement annoté a permis de mettre en évidence la démarche de l'expert face à la validation. Outre les caractéristiques linguistiques de la variation, l'expert va faire intervenir un autre critère plus pragmatique qui concerne la pertinence du terme par rapport au domaine étudié. Un terme va être rejeté s'il est jugé, soit trop générique, c'est-à-dire que son contenu informatif n'est pas assez riche (The Cell ← the adenoma cell), soit sans intérêt par rapport au domaine étudié (Medical record ← medical and pathology records).

Les quatre combinaisons de ces deux types de critères ont pu être observées dans notre corpus d'étude et sont résumées à l'aide du tableau suivant :

	Validité linguistique	**Pertinence du terme**
Comportement 1	OUI	OUI
Comportement 2	OUI	NON
Comportement 3	NON	NON
Comportement 4	NON	OUI

Tableau 4 : Comportement de l'expert face à la validation

En observant les trois premiers comportements, nous constatons que la validité linguistique d'une variation ne suffit pas à l'acceptation d'un terme. La pertinence semble constituer un critère décisif pour l'expert. Le

quatrième comportement de l'expert face au processus d'indexation confirme ce fait mais nous amène également à réfléchir plus généralement sur les fondements mêmes de l'indexation automatique, en tant que méthode d'extraction de connaissances. En premier lieu, face à ces résultats, il apparaît essentiel de préserver l'intervention de l'expert afin d'éliminer le bruit produit par le processus d'indexation automatique. En second lieu, l'acceptation d'un concept issu d'une mauvaise variation traduit clairement l'utilisation par l'expert de ses connaissances pour conserver de l'information implicite. Ses connaissances l'amènent à récupérer de l'information, qui semble pertinente par rapport au domaine d'étude, alors même que cette information n'est pas explicitement exprimée dans le texte. La question que nous pouvons nous poser est de savoir s'il faut tolérer cette part d'implicite, voire d'inférence, dans un processus d'indexation automatisée faisant appel à des traitements linguistiques tels que ceux mis en œuvre ici. En d'autres termes, ne devons-nous pas interdire, dans une approche d'extraction de connaissance reposant sur des techniques de TALN, que l'expert produise lui-même une part de cette connaissance ? En effet, ces techniques sont-elles réellement appropriées pour extraire de l'information implicite ?

L'analyse des termes, issus d'une mauvaise variation mais validés par l'expert, nous amène à prendre une position stricte sur ce sujet. Il est apparu que ces termes, bien qu'ils véhiculent la connaissance du domaine, n'expriment pas nécessairement la connaissance contenue dans le texte, même implicitement. Au contraire, dans de nombreux cas, le concept ramené par une mauvaise variation s'avère en contradiction avec le sens du texte. En conclusion, l'expert ne peut pas et ne doit pas s'appuyer sur des procédures de TALN pour extraire de l'information implicite. Ceci implique que les mauvaises variations doivent être systématiquement rejetées. Les modifications apportées aux métarègles vont dans ce sens. Il faut parvenir à filtrer au maximum les mauvaises variations pour augmenter la précision du système mais, également, pour empêcher l'expert d'acquérir de l'information implicite par ce biais.

4.2. Bilan sur l'analyse linguistique des séquences textuelles

Les termes (composés de deux ou trois unités) renvoient à des groupes nominaux plus ou moins complexes, dont la structure syntaxique exprime des relations de dépendance entre une tête et un autre élément (qu'il soit modifieur ou actant selon la distinction opérée dans le cadre de la grammaire de dépendance (Kahane, 2001)). L'analyse linguistique de l'ensemble des séquences textuelles a montré qu'une mauvaise variation agit au niveau syntaxique en modifiant les rapports de dépendance initiaux. L'altération de ces relations de dépendance se répercute au niveau sémantique en modifiant

le sens initialement véhiculé par le terme. Deux principaux phénomènes sont à l'origine de cette modification : (i) une mauvaise délimitation de la ST en corpus, et (ii) l'insertion d'une unité grammaticale et/ou lexicale dans le syntagme nominal.

*4.2.1. ERREUR DE DÉLIMITATION DE LA SÉQUENCE EN CORPUS***.** Les possibilités de transformation sur un terme sont exprimées dans les métarègles sous forme d'expressions régulières. Ici, toutes les métarègles reposent sur le même principe : les éléments qui délimitent les frontières du terme bornent l'expression renvoyant à la variation, tel que par exemple :

X2 N3 → **X2** C4 < {A|N|Np|V} 0-3 > **N3**

Dès lors que le système reconnaît dans la phrase le motif décrit par la métarègle, il extrait la séquence textuelle et renvoie au terme de référence. Ce qui signifie qu'il n'y a pas obligatoirement de correspondance syntaxique stricte entre la ST retrouvée et le syntagme nominal qu'elle est censée recouvrir en corpus. Les erreurs de découpage du syntagme en contexte constituent un des problèmes majeurs de la reconnaissance automatique des termes que les expressions régulières ne peuvent résoudre. Ces problèmes de circonscription du SN touchent à la fois les contextes droits et gauches.

Au niveau du contexte droit, une mauvaise délimitation de la ST en corpus peut altérer de deux manières les relations de dépendances initiales entre une tête et son expansion.

Dans le premier cas, l'unité, ayant la fonction de tête de syntagme dans la structure de départ, devient l'expansion d'une autre tête en corpus et perd en même temps toute relation avec son propre dépendant (le contexte récupéré à partir du corpus textuel est indiqué en italique dans les exemples ci-dessous) :

XX,1,Coor :
skeletal bone ↚ skeletal survey and bone *marrow examination*
XX,7,Coor :
T3 antibody ↚ t3, stsh and tsh receptor *antibody*
XX,10,Coor :
heart muscle ↚ heart, liver, kidney and skeletal muscle *tissue*
XX,28,Ins :
blood cell ↚ blood, tumor cell *lines*

Dans le second cas, les modifications syntaxiques touchent principalement l'expansion, qui se révèle en corpus dépendante d'une autre tête.

XX,34,Perm :
primary hypothyroidism ⇜ hypothyroidism (primary *myxedema*)
XX,31,Perm :
membrane protein ⇜ protein are membrane-*associated tranducers*
XXX,7, Ins :
biopsy of tongue ⇜ biopsy of the tongue *nodule*

Nous rencontrons des problèmes similaires lorsqu'il s'agit du contexte gauche. Dans ce cas, l'élément constituant la frontière gauche du terme est principalement affecté par des modifications de dépendance : au départ expansion dans le terme initial, il devient la tête d'un autre syntagme en corpus :

XX,4,Coor :
thymus gland ⇜ *rat* thymus and adrenal gland
XX,10,Coor :
thyroid tumors ⇜ *follicular* thyroid, breast, colon, and skin tumors
XX,7,Coor :
tumor cell ⇜ *pancreatic islet cell* tumor, paraganglioma, or merkel cell
XX,28,Ins :
disease markers ⇜ *autoimmune thyroid* disease, activation markers

À travers ces différents exemples, nous constatons qu'il est souvent difficile de juger de la validité d'une variation si nous n'avons pas accès au contexte plus large (tumor tissue ← tumor, normal tissue). Parfois, l'expert est en mesure de prendre une décision à partir des informations fournies par la ST lorsque celle-ci se révèle asémantique, telle que « skeletal survey and bone ». Toutefois, notre objectif est d'obtenir une meilleure reconnaissance de la variation par le système en introduisant des critères filtrants dans les métarègles.

4.2.2. DÉPENDANCES MODIFIÉES PAR INSERTION D'UNITÉS GRAMMATICALES ET/OU LEXICALES. D'autres facteurs sont susceptibles de modifier les dépendances. L'insertion d'une unité lexicale et/ou grammaticale (coordination, préposition) dans le syntagme de départ peut provoquer des ruptures syntaxiques évidentes pour un être humain. L'objectif est de parvenir à formaliser ces phénomènes linguistiques, afin que ces mauvaises variations soient automatiquement filtrées par le système.

(1) XX,1,Coor :

breast tissue	⇜	breast lesions or normal tissues
temperature receptor	⇜	temperature dependent and receptor

(2) XX,37,Perm :

thyroid function	⇜	function in a rat thyroid

Tout en préservant les bonnes variations :

(3) XX,1,Coor :

dividing cell	←	dividing follicular and stroma cells
Thyroid function	←	thyroid growth and function

(4) XX,37,Perm :

thyroid tumor	←	tumor of the thyroid

XX,37,Perm :

animal tissue	←	tissues in animal

L'insertion d'une coordination et d'un substantif à sa droite peut éclater le groupe nominal initial en deux groupes totalement indépendants (1). L'insertion d'une préposition risque de modifier ou d'introduire des relations actancielles entre les éléments qui ne respectent pas le sens initial (2). Dans le premier cas, il faut autoriser soit l'introduction d'un adjectif à gauche de la coordination, suivie de l'insertion de n'importe quelle unité à sa droite, soit l'introduction d'un substantif sans autre insertion (3). Dans le second cas, l'introduction de la préposition *of* garantit la préservation du sens initial (4), alors que la préposition *in* n'est sémantiquement acceptable que si le terme initial exprime une relation méronymique ou locative entre la tête et l'expansion.

5. DÉFINITION DE CRITÈRES SYNTAXIQUES ET MORPHO-SYNTAXIQUES

Notre objectif est de parvenir à une meilleure reconnaissance de la variation syntaxique par le système en obtenant des métarègles plus filtrantes. Nous cherchons à intervenir sur ces métarègles afin qu'elles soient en mesure de préserver les rapports de dépendance, entre têtes et expansions, lors des transformations linguistiques. L'analyse linguistique des ST de notre corpus nous amène à définir deux types de critères qui vont contraindre les métarègles à ne pas disloquer les dépendances : le critère syntaxique, par la modification des expressions régulières, et deux critères morpho-dérivationnels, qui sont exploités au niveau des expressions régulières et des conditions qui sont associées à chaque métarègle.

Le critère syntaxique permet d'interdire certaines structures qui éclatent obligatoirement les relations initiales. Les critères morphologiques sont flexionnels par la marque du pluriel sur les noms têtes (notamment lors de

coordination) et grammaticaux par l'appartenance à une partie du discours spécifique.

Étant donné les propriétés linguistiques propres à chaque type de transformation, il est impossible d'appliquer ces critères uniformément à toutes les métarègles, quelque soit leur type. Ainsi, les transformations par coordination sont particulièrement propices à ce genre de manipulation, alors que les transformations par insertion sont beaucoup plus problématiques de ce point de vue. D'autre part, les critères sont exploités afin d'assurer que les relations de dépendances soient préservées de manière interne – *i.e.* au niveau de la ST – et de manière externe – *i.e.* au niveau de la délimitation de la ST. Une métarègle exploitant l'ensemble de ces paramètres sera beaucoup plus filtrante et précise qu'une métarègle qui ne les introduit que partiellement. Pourtant, celle-ci est tout à fait en mesure de produire de bonnes variations telles que : thyroid vein ← thyroid artery and vein. C'est pourquoi, il faut préserver ces métarègles moins filtrantes afin de ne pas pénaliser le rappel sur la précision. D'autre part, dans de tel cas, seul le contexte est susceptible de lever l'ambiguïté syntaxique, ainsi que de confirmer ou infirmer la validité de la variation. Il se révèle donc indispensable de fournir à l'expert une fenêtre de décision plus large que celle correspondant à la ST.

Nous illustrons ci-dessous ces différents cas à partir de l'exemple concret de deux métarègles, l'une de coordination et l'autre d'insertion.

5.1. Application des critères sur une métarègle de coordination

Considérons cette première métarègle

> **« X2 N3 → X2 < {A|N|Np|V} 1-3 PUNC/','? > C4 < {A|N|Np|V} ? > N3 »** (XX, Coor, 1, tableau 2)

générant plusieurs structures variationnelles incorrectes (5 à 7), correctes (8-9) ou ambiguës (10). Dans ce dernier cas, seul le contexte peut lever l'ambiguïté.

(5) Breast tissue ↚ breast lesions or normal tissues
(6) Tumor cells ↚ tumor patterns or cell
(7) Skeletal tissue ↚ skeletal muscle, and lymphoid tissue
(8) Dividing cell ← dividing follicular and stroma cells
(9) Endocrine cell ← endocrine tissues and cells.
(10) ? Thyroid vein ← thyroid artery and vein

Partant de ces observations, la métarègle initiale a été modifiée au niveau syntaxique et morphologique, à partir de critères flexionnels et grammaticaux tels que l'appartenance à une catégorie lexicale spécifique. Introduites en cascade, ces modifications vont permettre de ne ramener que des variations correctes ou ambiguës, et de répartir ces variations selon

plusieurs degrés de confiance. Dans tous les cas, nous avons interdit l'introduction d'une ponctuation. Tous les exemples de telles structures rencontrés dans notre corpus d'étude ont montré que la présence d'une ponctuation rompt les relations de dépendances initiales.

1ère modification : X2 N3 → X2 < {A|N|Np|V} 0-2 > A4 C5 < {A|N|Np|V} > N3
Dividing cell ← dividing follicular and stroma cells

Cette première « sous-métarègle » autorise uniquement l'insertion d'un adjectif à la gauche de la coordination C5. En raison des propriétés syntaxiques de l'anglais, une telle structure implique obligatoirement que A4 soit rattaché à N3, ce qui permet l'introduction de n'importe quelle unité lexicale à la droite de cette coordination. La coordination s'applique aux expansions de la tête N3. La métarègle génère ainsi des ST dont les rapports de dépendance ne sont ni ambigus, ni modifiés.

2de modification : X2 N3 → X2 < {A|N|Np|V} 0-2 > N4 C5 N3
<N4 agr num> = plu
<N3 agr num> = plu
Endocrine cell ← endocrine tissues and cells.

Lorsque l'unité lexicale introduite à gauche de C5 renvoie à un substantif, il faut interdire l'insertion de toute unité à droite de la coordination sous peine d'éclater les dépendances initiales ((1) Breast tissue ↚ breast lesions or normal tissues). Ce type de métarègle exprime une coordination sur les têtes du syntagme : N4 et N3 partagent les mêmes expansions. L'ambiguïté liée à la délimitation de la ST en corpus (contexte droit) peut être levée en introduisant un critère flexionnel. Après consultation du corpus, il est apparu que la marque du pluriel sur N4 et N3 indiquait l'appartenance des deux substantifs au même SN, respectant ainsi les relations de dépendances initiales entre N3 et son expansion X2.

3ème modification : X2 N3 → X2 < {A|N|Np|V} 0-2 > N4 C5 N3
<N4 agr num> ! plu
<N3 agr num> ! plu
Thyroid vein ← thyroid artery and vein

Enfin, certaines variations, bien qu'ambiguës, se sont révélées correctes après consultation du contexte plus large en corpus. Par rapport aux deux métarègles précédentes, cette dernière n'est filtrante que par rapport au critère syntaxique. En l'état, son manque de précision nécessite obligatoirement une intervention humaine. De plus, il est indispensable de fournir d'autres critères d'aide à la décision qui soient susceptibles de lever

l'ambiguïté sur les rapports de dépendance entre N3 et X2. Ces informations ne peuvent être apportées que par les contextes gauche et droit de la ST. Nous constatons dès lors les limites de nos critères dont la portée se restreint à l'expression variante. Dans un premier temps, nous pallierons ce problème en permettant une visualisation plus large du contexte de la ST pour validation humaine. Mais notre objectif est de parvenir à une méthode qui permettrait de s'affranchir des frontières imposées par le terme de départ et de vérifier automatiquement le correct découpage de la séquence en corpus.

5.2. Application des critères sur une métarègle d'insertion

Il est apparu plus problématique d'appliquer nos critères internes sur les métarègles d'insertion. Particulièrement en ce qui concerne la métarègle initiale *X2 N3 → X2 < {A|N|Np|V} 0-3 > N3* :

(11) Cultured Cell cells	? ←	cultured neoplastic human thyroid
(12) Growth growth	? ←	decreased tumor cell growth
(13) Tumor cells	↚	tumor growth creates cells
(14) Cell differentiation	↚	cell line exhibiting differentiation
(15) Thyroid tumor tumors	? ←	thyroid are unrelated malignant
(16) Tumor antigen	? ←	tumor associated antigen.

Aucun critère morphologique ne peut être appliqué afin d'assurer la correcte délimitation de la séquence en corpus. Nous ne pouvons intervenir qu'au niveau de la syntaxe en différenciant les transformations, avec et sans l'introduction d'un verbe (à la forme conjuguée, participative ou gérondive). De plus, nous contraignons l'insertion du verbe immédiatement après N2 afin de préserver des structures qui peuvent se révéler correctes (15-16) et interdire certaines autres (13-14).

1ère modification :	X2 N3 → X2 < {A|N|Np} 0-3 > N3
2de modification :	X2 N3 →X2 V4 < {A|N|Np} 0-2 > N3

Dans ces deux cas, les formes variantes nécessitent de se reporter au contexte plus large. Les modifications n'apportent qu'une aide à la personne chargée de la validation.

Ces deux exemples nous ont permis d'illustrer les disparités dans l'application de nos critères linguistiques au niveau des métarègles, et de la nécessité d'aller plus loin dans notre démarche pour être en mesure d'intervenir sur le contexte syntaxique externe aux ST ramenées par le système.

6. CONCLUSION

Nous avons tâché de montrer comment il était possible, à partir d'une analyse linguistique des séquences textuelles issues d'une variation, de définir un certain nombre de critères syntaxiques et morpho-syntaxiques pour rendre les métarègles plus filtrantes. Ces critères contraignent les transformations à préserver les relations de dépendance initiales entre têtes et expansions afin de conserver le sens véhiculé par les termes. Ces dépendances pouvant être altérées, soit par introduction d'unités linguistiques ou de ponctuation, soit par une mauvaise délimitation des ST en corpus. Toutefois, nous avons également montré que ces critères ne sont pas applicables de manière homogène à toutes les métarègles et ne permettent pas d'obtenir dans tous les cas des métarègles totalement filtrantes. Ce constat révèle les limites de notre approche : il faut pouvoir contourner l'impossibilité d'appliquer les critères internes à certaines métarègles en utilisant des indices externes qui permettraient de contrôler la correcte délimitation des ST en corpus. Nous souhaitons explorer cette voie en testant la possibilité d'introduire, dans les traitements mis en œuvre lors de la reconnaissance de la variation, une analyse syntaxique partielle type *chunking* qui permettrait de définir les frontières des syntagmes nominaux en corpus et, ainsi, de vérifier l'identité syntaxique entre la ST extraite et le segment nominal en corpus. L'augmentation du pouvoir filtrant des métarègles se révèle indispensable pour obtenir une indexation linguistiquement fiable. Elle permet parallèlement de réduire au maximum les possibilités de préserver de l'information implicite à partir de techniques TALN.

7. RÉFÉRENCES

Basili R.; Moschitti A.; Pazienza M.-T.; Zansotto F. B. 2001. "A constrative approach to term extraction", *Actes des 4èmes Rencontres Terminologie et Intelligence Artificielle TIA'01*, Nancy 3-4 mai 2001, p. 119-128.

Bourigault D.; Jacquemin C. 2000. "Construction de ressources terminologiques", J.-M. Pierrel (ed), *Ingénierie des langues*, Paris : Hermes, p. 215-233.

Bourigault D.; Jacquemin C.; L'homme M.-C. 2001. *Recent Advances in Computational Terminology*, Natural Language Processing, Amsterdam/Philadelphia : John Benjamins Publishing Company.

Daille B. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, Thèse en information fondamentale, Université de Paris VII.

Daille B.; Habert B.; Jacquemin C.; Royaute J. 1996. "Empirical observation of term variation and principles for their description", *Terminology*, vol. 9, n° 2, p. 171-216.

David S.; Plante P. 1990. "De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes", *Intelligence Artificielle et Sciences Cognitives au Québec*, 3(3), p. 140-154.

Enguehard C. 1992. *Acquisition naturelle automatique d'un réseau sémantique*, Thèse de doctorat de l'Université de Technologie de Compiègne, décembre 1992.

Ibekwe-Sanjuan F. 1998. "Terminological variation, a means of identifying research topics from texts", *Proceedings of the Joint International Conference on Computational Linguistics (COLING-ACL'98)*, Montréal Québec, 10-14, August 1998, p. 564-570.

Jacquemin C. 1997. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*, Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.

Kahane S. 2001. "Grammaires de dépendance formelles et théories Sens-Texte", *Actes de la 8ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles, TALN'01, tutoriel*, p. 17-76.

Morin E.; Jacquemin C. 1999. "Expansion automatique de thesaurus à partir de corpus", *Actes de la Troisième Conférence sur l'Ingénierie des Connaissances (IC'99)*, Palaiseau, France, Juin 1999, p. 97-105.

Polanco X.; François C. 2000. "Data Clustering and Cluster Mapping or Visualization in Text Processing and Mining", *Sixth International ISKO Conference*, Toronto, Canada, Advances in Knowledge Organization, Vol. 7, p. 359-365.

Pozzi M. 2002. "Towards the harmonisation of terminology of ISO/TC37 standards : corpus-based identification of problem terms", *Proceedings of the 6th International Conference of the Terminology and Knowledge Engineering, TKE'02*, p. 101-106.

Rastier F. 1995. "Le terme : entre ontologie et linguistique", *Actes des 1ères Journées TIA*, Villetaneuse, La banque des mots, Numéro spécial 7-1995, p. 35-65.

Rinaldi F.; Dowdall J.; Hess M.; Kaljurand K.; Koit M.; Vider K.; Kahusk N. 2002. "Terminology as knowledge in answer extraction", *Proceedings of the 6th International Conference of the Terminology and Knowledge Engineering, TKE'02*, p. 107-112.

Royaute J. 1999. *Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information*, Université Henri Poincaré Nancy I. Thèse de doctorat en informatique.

Royaute J.; François C.; Zasadzinski A.; Besagni D.; Dessen P.; Maunoury M. T.; Le Minor S. 2004. "Relation entre gènes impliqué dans les cancers de la thyroïde", *Revue des Nouvelles Technologies de l'Information (RNTI-E-2)*, EGC 2004, vol.II, p. 465-476.

Salton G. 1986. "Another look at automatic text-retrieval systems", *ACM (Association for Computing Machinery)*, vol. 29, n° 1, p. 648-656.

Smadja F. 1993. "Retrieving collocations from texts : Xtract", *Computational linguistics*, 19(1), p. 143-177.

Schmid H. 1994. "Probabilistic part-of-speech tagging using decision trees", *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, p. 44-49.

8. NOTES

(1) L'étude que nous vous présentons s'inscrit dans un processus de fouille de textes mené dans le cadre d'un Projet inter-EPST, soutenu par l'INSERM et en partenariat avec l'Institut Gustave Roussy de Villejuif (Royauté & al. 2003 ; Zasadzinski 2002). Appel d'offre 2000 « Bioinformatique inter-EPST » (CNRS, INRA, INRIA, INSERM).

(2) CELEX est une base de données lexicales conçue par le « Centre of Lexical Information, Max Plank Instutitute for Psycholinguistics, Nijmegen, Dutch » et en libre accès (http://www.kun.nl/celex/).

(3) L'UMLS est un projet de l'U.S. Department of Health and Human Services, National Institutes of Health (NIH) – National Library of Medicine (NLM). Elle constitue la ressource la plus importante et la plus complète dans le domaine de la biologie et de la médecine.

LES BESOINS D'INTERACTIONS EN TRAITEMENT AUTOMATIQUE DES LANGUES ET EN LINGUISTIQUE DE CORPUS : ÉTUDE DE CAS

Stéphane Ferrari, Vincent Perlerin
Université de Caen − GREYC, CNRS UMR 6072

1. INTRODUCTION

Dans cet article, nous cherchons à montrer, par le biais d'exemples, en quoi certaines pratiques de Traitement Automatique des Langues (TAL) et de linguistique de corpus font émerger des besoins d'interaction entre les utilisateurs et les machines. Nous illustrons nos propos par l'analyse de divers travaux fondés sur un même modèle linguistique. Il apparaît ainsi que, parmi des tâches aussi éloignées que la construction de ressources lexicales et l'analyse de résultats de traitements automatiques sur corpus, des régularités peuvent se dégager en ce qui concerne les interactions mises en jeu. Dans une période où renaît la réflexion sur l'instrumentation de la linguistique, nous proposons un regard sur la nature de ces régularités, en interrogeant les relations entre modèles et outils informatiques.

Nous présentons dans une première partie le cadre de nos expériences en TAL et en linguistique de corpus, et plus spécifiquement le modèle LUCIA sur lequel se fondent nos travaux. Ce modèle permet de représenter et d'organiser des connaissances lexico-sémantiques pour les exploiter ensuite lors d'analyses instrumentées de documents textuels.

Après avoir exposé les grandes lignes de notre approche, la deuxième partie de cet article est plus spécifiquement consacrée à l'analyse des interactions mises en jeu lors de la phase de constitution des ressources lexicales. Nous appuyons principalement notre réflexion sur l'utilisation du logiciel d'étude LUCIABUILDER, développé par nos soins pour permettre l'organisation effective de lexiques sémantiques selon les principes du modèle LUCIA. En s'intéressant plus aux aspects fonctionnels de cette phase, un parallèle peut être établi avec les outils informatiques classiques de gestion de données. Cependant, pour une meilleure interaction, des besoins spécifiques émergent en ce qui concerne les représentations visuelles des données et leur manipulation. Ces besoins sont motivés par la nécessité de

permettre à l'utilisateur de prendre en main le modèle de la langue que les outils mettent en œuvre.

Dans une troisième partie, nous proposons une présentation comparée de deux applications permettant l'analyse automatique d'une collection de documents. L'une est destinée à des experts du modèle étudiant un fait de langue, l'autre peut être utilisée par un utilisateur novice pour une recherche documentaire sur ses centres d'intérêts personnels. Dans les deux cas, ces applications doivent être en mesure de fournir des résultats facilement interprétables et rapidement exploitables. Une étude de ces contraintes communes permet de préciser les représentations visuelles et les interactions à mettre en œuvre dans les outils qui instrumentent nos modèles. Nous précisons ainsi des besoins relatifs à la navigation dans une collection de documents, et d'autres relatifs à la représentation à différentes échelles d'un même document, pour un repérage rapide ou une analyse approfondie. Nous analysons le caractère générique de ces besoins et leur dépendance éventuelle vis-à-vis du modèle, de la tâche et de l'utilisateur.

En conclusion, nous revenons sur l'étude de cas présentée et ouvrons notre réflexion en questionnant la pertinence de standards pour la représentation, l'échange de données ou encore pour l'intégration d'outils au sein de plates-formes d'ingénierie linguistique.

2. CADRE DES EXPÉRIENCES, MODÈLE SOUS-JACENT

Les besoins d'interactions spécifiques ou de représentations visuelles particulières que nous discutons dans cet article émergent avant tout de nos propres expériences. C'est pourquoi nous présentons dans cette section les grandes lignes du modèle LUCIA, sur lequel se fondent nos travaux de recherche, à la croisée du TAL et de la linguistique de corpus.

Nous précisons d'abord la manière dont les ressources sont structurées pour rendre compte d'éléments de sens communs, ou proches, entre des entrées lexicales. Nous présentons ensuite le principe d'analyse automatique mis en œuvre pour fournir une aide à l'interprétation de documents textuels.

3. RESSOURCES LEXICO-SÉMANTIQUES

Le modèle, LUCIA, s'inspire d'une lignée de travaux autour du modèle ANADIA (Coursil 1992 ; Beust 1998), et de la *Sémantique Interprétative* (Rastier, 1987). En tant que modèle de représentation lexicale, LUCIA permet de décrire des éléments de connaissance propres à un utilisateur ou à un groupe d'utilisateurs partageant, dans le cadre déterminé d'une tâche commune, une même vision sur le lexique d'un domaine.

La description des entrées lexicales se fait de façon componentielle à travers l'utilisation de la notion de sème telle qu'introduite dans Beust (1998). Le sème est ici envisagé comme un attribut constitué d'un jeu d'oppositions de valeurs. Cette notion s'éloigne légèrement de la notion classique pour une raison essentielle d'opérabilité : elle permet en effet de définir des critères organisationnels, qui tiennent compte, à la fois, des points communs et des différences entre les entrées lexicales. Par exemple, les lexies *anticyclone* et *dépression* peuvent être toutes deux en partie décrites à l'aide de l'attribut [Pression : basse *vs* haute], *anticyclone* actualisant la valeur « haute » tandis que *dépression* actualise la valeur « basse ». Cette opposition locale permet de différencier finement les deux lexies.

Plusieurs attributs peuvent par ailleurs être combinés pour décrire un ensemble de lexies proches. Les lexies décrites par un jeu d'attributs communs peuvent être regroupées dans une structure appelée *table*, dont chaque ligne correspond à une actualisation spécifique des valeurs mises en jeu. Il est ainsi possible (voir figure 1) de regrouper dans une même table des *« Phénomènes météorologiques dynamiques »* les lexies *vent*, *accalmie*, *assombrir*, *éclaircie*, *réchauffement*, *fonte*, *tempête de neige*, …, en les décrivant localement avec des actualisations différentes des deux attributs [Axe : agitation *vs* température] et [Direction : monte *vs* descend].

Enfin, une notion d'héritage sémique peut s'exprimer par un lien orienté d'une ligne vers une table. Ainsi, dans l'exemple de la figure 1, la lexie *vent* est décrite sur une ligne de la table précédemment proposée. Dans une autre table héritant de cette ligne (*i.e.* des actualisations des valeurs des attributs décrivant la lexie *vent*), des lexies employées pour parler de vents sont différenciées selon leur zone géographique.

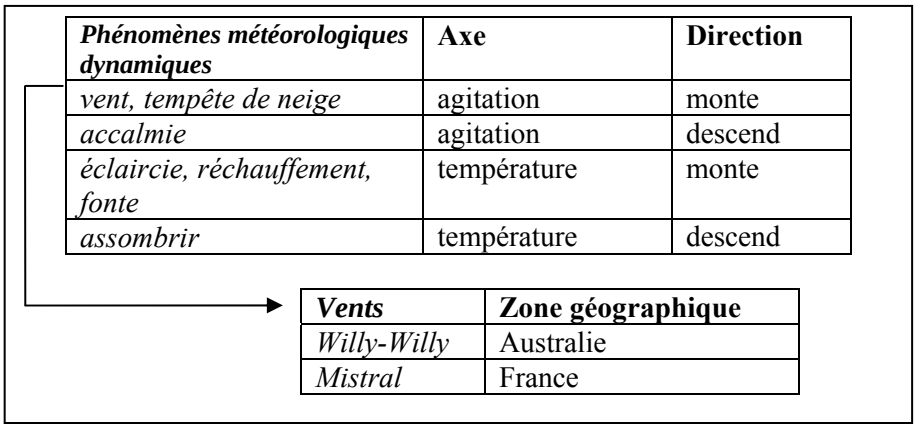


Figure 1 : Héritage sémique et lien de ligne à table

Un ensemble de tables, ainsi reliées et décrivant un domaine particulier, est appelé un dispositif. Nous renvoyons à Perlerin *et al.* (2002) pour plus de détails sur ces notions qui seront illustrées dans la section 0, où nous présentons plus spécifiquement l'interface développée pour la construction interactive de telles ressources. La pertinence des descriptions obtenues avec ce modèle, leur cohérence et leur justesse sont à la discrétion de l'utilisateur. On peut ainsi questionner, dans l'exemple précédent, la description de la lexie *tempête de neige* : une « baisse de température » peut tout autant décrire cette lexie qu'une « montée en agitation », ce qui d'une certaine manière interroge le choix du jeu d'oppositions initial des « axes » agitation et température. De telles remises en question font partie du modèle centré utilisateur : il a été montré dans Perlerin *et al.* (2002) comment intégrer, au cycle d'utilisation du modèle, un processus de révision des ressources lexicales. Nous considérons, pour la suite de cet article, que les ressources utilisées en illustration s'inscrivent dans une telle démarche : elles sont susceptibles d'être révisées après utilisation et ne sont que des instantanés sortis d'un cycle d'expérimentations.

4. AIDE À L'INTERPRÉTATION

Le modèle LUCIA possède un second volet qui concerne l'interprétation. En tant que modèle de l'interprétation, il propose d'exploiter des connaissances lexico-sémantiques, décrites comme indiqué en 3, pour mieux appréhender le matériau textuel. Le principe des analyses proposées s'inspire de la notion d'isotopie, envisagée ici comme la redondance d'un attribut, ou d'une valeur d'attribut, dans une unité textuelle.

Ainsi, l'exemple suivant contient bien une redondance de l'attribut [Pression : basse *vs* haute] même si chaque lexie n'en actualise pas la même valeur :

E1 « S'il s'agit d'un anticyclone éphémère entre 2 passages de dépressions (dorsale), l'air est en général un peu plus frais et porteur d'une instabilité un peu plus marquée. ».[1]

La phase d'analyse automatique des textes consiste donc essentiellement à déterminer quelles redondances existent, dans quel type d'unité (paragraphe, texte, collection), en projetant pour chaque occurrence d'une lexie l'ensemble des sèmes qui la décrivent dans les ressources qu'un utilisateur aura construites. Mais l'aide que le modèle fournit pour l'interprétation des documents s'inscrit avant tout dans l'interaction avec l'utilisateur, et dans les représentations qui lui sont proposées des résultats de cette analyse automatique. La redondance sémique n'est pas une donnée suffisante pour conclure, elle constitue plutôt un bon aiguillage vers une interprétation qui doit, au final, être menée par l'utilisateur.

Dans nos travaux, nous avons plus particulièrement utilisé le modèle pour deux applications assez éloignées : le projet ISOMETA visant à l'analyse sur corpus d'une métaphore conceptuelle, d'une part, et l'aide à la recherche documentaire, d'autre part. Il est rapidement apparu que les résultats d'analyse ne sont pas à exploiter ni à présenter de la même manière selon les cadres applicatifs. Ainsi, pour étudier, dans une collection d'articles d'économie, la métaphore de la météorologie boursière, nous avons d'abord cherché à décrire assez finement le lexique des deux domaines source et cible de cette métaphore, *la Météorologie* et *la Bourse*, afin de faire ressortir plus précisément des redondances d'attributs communs aux deux domaines et accompagnant certains emplois métaphoriques. En revanche, lorsqu'il s'agit de recherche documentaire, nous sommes en général plutôt intéressés par l'ensemble des redondances relatives à un même domaine, le thème de la recherche, et la plupart du temps sans distinction particulière entre les attributs décrivant les lexies de ce domaine.

Considérons l'exemple suivant, extrait du corpus étudié dans le projet ISOMETA :

> *E2 Jeudi, changement de décor. Les nuées commencèrent à se dissiper, et quelques rayons de soleil pénétrèrent jusqu'au cœur de la Bourse. Las ! Ce ne fut qu'une embellie.*[2]

Pour le projet ISOMETA (Perlerin *et al.*, 2002 et Beust *et al.*, 2003), les redondances surlignées dans l'exemple E2, concernant les lexies *nuées*, *rayons de soleil* et *embellie,* sont à présenter à l'utilisateur de manière particulière, car elles relèvent du domaine source de la métaphore étudiée, *la Météorologie*. Il est à noter que de nombreuses autres redondances existent dans le même document, concernant cette fois le domaine cible, *la Bourse*. Le modèle ne permet pas de décider de la présence de métaphores, ce n'est d'ailleurs pas sa vocation. Mais il doit être en mesure de fournir à l'utilisateur des informations pertinentes pour l'aider à repérer et à interpréter des emplois comme ceux de l'exemple E2.

Supposons maintenant que, pour une autre étude, un utilisateur recherche des documents traitant de météorologie. Celui dont est extrait l'exemple E2 sera plutôt à écarter des réponses à lui proposer car, bien que des lexies en rapport avec le domaine recherché soient présentes, elles y sont en quantité restreinte, et les redondances mises en jeu ont une portée très limitée dans le texte. Une fois encore, le modèle n'a pas pour but de décider de l'inadéquation du document à la recherche de l'utilisateur, mais il doit être capable de présenter à l'utilisateur qu'une « trace » de lexique météorologique y existe, en lui donnant les moyens de juger rapidement de sa pertinence.

L'aide à l'interprétation dépend donc nettement de l'application, notamment lorsqu'on l'envisage du point de vue de l'interaction. La section

8 est consacrée à une présentation plus détaillée des moyens mis en œuvre pour adapter les représentations visuelles et les interactions pour les deux applications présentées, de manière à satisfaire les objectifs qui viennent d'être mis en avant pour chacune d'entre elles. Nous commençons par présenter, dans la section 5, la façon dont s'opère la phase commune à ces deux applications, celle qui concerne la constitution des ressources.

5. INTERACTIONS POUR LA CONSTITUTION DES RESSOURCES

Le modèle LUCIA repose sur quelques principes simples de description sémantique du lexique et d'analyse des redondances d'attributs. Ceci étant, malgré leur simplicité, ces principes ne sont pas nécessairement connus d'un utilisateur novice. Or, un des objectifs de nos travaux consiste à proposer une aide à des utilisateurs novices, pour les tâches qu'ils peuvent être amenés à réaliser sur des documents et qui nécessitent une interprétation du matériau textuel. C'est pourquoi nous présentons dans cette section le logiciel d'étude que nous avons développé pour permettre la spécification de ressources lexicales selon les critères du modèle LUCIA. Nous commençons par une comparaison des fonctionnalités attendues avec celles offertes par d'autres outils informatiques à visée plus générique, en expliquant en quoi ces outils ne sont pas satisfaisants ici. Nous montrons ensuite quelles interfaces nous utilisons pour commencer à apporter une solution aux problèmes non résolus par les outils classiques.

6. DES FONCTIONNALITÉS À DEUX NIVEAUX

Les ressources lexicales utilisées pour l'analyse des documents peuvent être décrites en termes de structure et de données. Dans le modèle LUCIA, les attributs constituent la partie structurante, leurs combinaisons déterminant à la fois les regroupements en tables et les liens d'héritage. Les lexies constituent quant à elles les données que décrit la partie structurante. Pour l'étape de constitution des ressources, les premières fonctionnalités qui émergent d'une analyse des besoins sont donc relativement classiques :

- créer, modifier, supprimer des structures ;
- ajouter, modifier, supprimer des données.

Si l'on raisonne à ce stade en termes de modèle informatique pour mettre en œuvre le modèle de TAL, les solutions qui semblent s'imposer consisteraient à choisir un modèle de représentation informatique et d'exploiter des outils déjà existants pour les fonctionnalités évoquées. Les systèmes de gestion de bases de données (SGBD) classiques permettraient de remplir ces fonctions avec efficacité. Cependant, les utilisations de notre modèle vont un peu à l'opposé des opportunités offertes par un SGBD,

notamment en ce qui concerne l'utilisation de données en grand nombre ou de manière concurrentielle. En plaçant l'utilisateur au cœur du modèle, il semble qu'il vaille mieux envisager des petites bases individuelles plutôt que de grandes bases partagées. L'aspect textuel de nos ressources peut aussi nous orienter vers le format XML. Dans cette optique, de nombreux outils existent qui permettraient de remplir les fonctionnalités précédentes. Cette solution ne présente pas d'inconvénient particulier, et le format XML est effectivement celui que nous utilisons pour le stockage de nos ressources. En revanche, les outils génériques pour l'édition de documents XML ne peuvent satisfaire l'ensemble des besoins inhérents à notre modèle. Une fonctionnalité supplémentaire émerge que ces outils ne couvrent plus : familiariser l'utilisateur avec le modèle.

Dans un premier temps, cette fonctionnalité peut être vue plus comme une contrainte, pour la réalisation de l'interface entre l'utilisateur et les ressources, que comme une véritable fonctionnalité. Ainsi, plutôt que d'attendre de l'utilisateur qu'il s'adapte au modèle de représentation informatique codant nos ressources, cette contrainte impose de créer une interface qui l'en dispense, en collant au modèle lexical plus qu'au modèle informatique. Mais la contrainte va en réalité plus loin dans notre approche : il s'agit non seulement de permettre à l'utilisateur de s'adapter au modèle de TAL, mais aussi et surtout de se familiariser avec ses propres ressources, en lui en offrant des points de vue multiples et en le préparant ainsi à la phase suivante d'exploitation des ressources. Comme il sera vu dans la section 8, la visualisation des résultats lors de la phase d'exploitation amènent finalement à prendre en considération une dimension graphique dès la constitution des ressources, en plus du modèle lexical. Cet aspect est absent de la présentation du modèle en 3 parce qu'il dépend essentiellement d'un choix d'interaction pour l'aide à l'interprétation. Nous montrons comment nous l'avons intégré à notre interface dans la suite de cette section.

7. LOGICIEL D'ÉTUDE *LUCIABUILDER*

Les besoins de créer des structures, d'y insérer des données, puis de modifier tant les structures que les données au fur et à mesure des usages peuvent être satisfaits par une interface indépendante du choix des représentations informatiques des données. Bien qu'ayant choisi le standard XML pour le stockage et l'échange, nous avons en l'occurrence fait d'autres choix techniques pour la représentation interne dans nos interfaces et nous n'exploitons pas directement l'ensemble des possibilités offertes par la sphère des technologies XML, en particulier les arbres DOM. Ainsi, pour mettre en œuvre les fonctionnalités évoquées plus haut, bien que développé en *Java*, le logiciel d'étude LUCIABUILDER utilise une représentation interne des données adaptée au modèle LUCIA. Il exploite en outre la librairie de

composants *Swing* pour l'interface graphique dont des copies d'écran sont proposées en illustration dans cette section.

L'interface se décompose en panels d'interaction dont certains offrent des fonctionnalités correspondant à des étapes obligatoires pour la création des ressources et ne peuvent donc être ignorés par l'utilisateur. En se référant à la figure 2, des onglets permettent d'accéder aux panels, les cinq visibles sur la figure correspondent aux fonctionnalités suivantes :

- « Attributs » : création, modification, suppression d'attributs ;
- « Tables » : ajout, modification, suppression d'entrées lexicales dans les tables ;
- « Topiques » : point de vue différentiel sur les données d'une même table ;
- « Dispositifs » : création des liens d'héritage ;
- « DictLex » : association de formes graphiques aux entrées lexicales, exploitant en outre la base de données lexicales MHATLEX de l'IRIT.

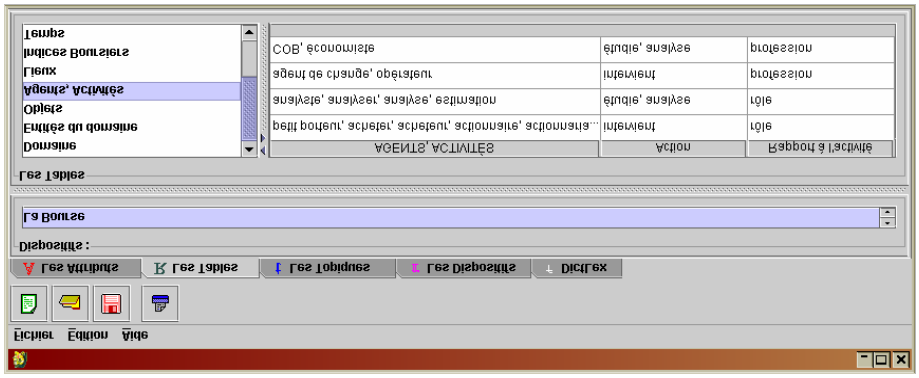


Figure 2 : Construction de tables

Dans le dispositif « La Bourse », les attributs [Action] et [Rapport à l'activité] permettent ici de différencier le lexique des acteurs boursiers : *petit porteur*, *analyste*, *agent de change* et *économiste*.

Le panel « Attributs » est utilisé pour la définition obligatoire des éléments de structure qui permettent d'organiser les ressources. Le panel « Tables » est tout autant nécessaire pour préciser quelles sont les lexies décrites par les éléments de structure. Pour la création d'une table, l'utilisateur précise quels attributs sont à utiliser localement, et le logiciel calcule automatiquement les combinaisons des valeurs d'attributs pour offrir une représentation en lignes et en colonnes (figure 2). À ce stade, il est possible qu'un nombre de lignes élevé perturbe l'appréciation des oppositions locales. C'est pourquoi le panel « Topiques » propose un point de vue d'une autre nature sur les tables (figure 3). Dans la représentation en

topique, chaque ligne d'une table est représentée par un rectangle contenant le premier mot de la ligne, et chaque rectangle est relié aux autres par des arcs indiquant quels attributs diffèrent dans les descriptions des lignes correspondantes. Lorsqu'une ligne ne contient aucune entrée lexicale, le rectangle qui la représente contient les valeurs des attributs la décrivant. Lorsqu'une table est construite à partir de nombreux attributs, le graphe ainsi obtenu peut être relativement complexe et difficile à lire. Il est de ce fait possible de préciser un nombre maximal de différences requises pour tracer les arcs entre rectangles. Dans la figure 3, le graphe tracé est contraint à *une* différence près, les arcs dessinés sont donc étiquetés par *un* attribut au maximum.

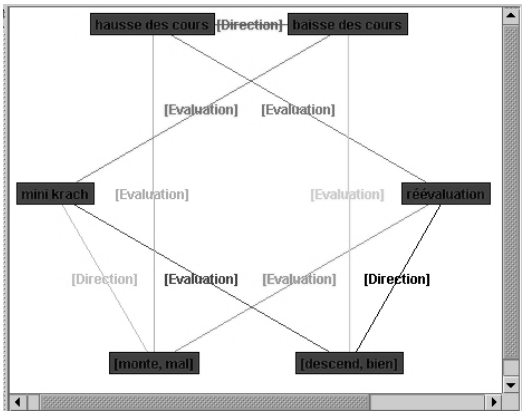


Figure 3 : Représentation en topique à **une** *différence près*

Pour les « Phénomènes dynamiques » de « La Bourse », l'utilisateur a différencié entre autre « baisse des cours » et « hausse des cours » par une actualisation différente de l'attribut [Direction].

Les représentations en topiques permettent d'apprécier la structure différentielle engendrée par les attributs considérés. Dans le cadre d'une création de ressources lexicales par un utilisateur novice, elles lui sont particulièrement utiles pour vérifier la cohérence des descriptions engendrées par les combinaisons de valeurs d'attributs, à l'origine de chaque table. L'utilisation conjointe d'un étiquetage des arcs par les attributs et des entrées lexicales (lorsque c'est possible) permet selon nous une familiarisation avec différentes notions :

- l'utilisateur apprend à s'abstraire du contenu lexical de chaque ligne pour n'en retenir éventuellement qu'un représentant unique ;
- il se familiarise avec les noms des attributs, forcément redondants dans ce graphe ;

– il s'abstrait aussi du nom des valeurs opposées, qui lui sont cachées, pour être remplacées par les lexies qu'elles décrivent sur chaque ligne.

L'ensemble de ces opérations amène finalement l'utilisateur à oublier, au moins temporairement, les noms qu'il a dû employer pour créer certains éléments de structure et ne plus en voir qu'un résultat local pour sa description lexicale : constater par exemple que les lexies *vent* et *accalmie* s'opposent localement de la même manière que les lexies *éclaircie* et *assombrir*, selon un attribut commun [Direction]. C'est ce genre de recul sur le modèle que nous cherchons à obtenir à travers les différentes représentations graphiques proposées dans l'interface de LUCIABUILDER. L'objectif est de faire en sorte que l'utilisateur puisse s'approprier les notions sur lesquelles se fonde le modèle, en exprimant son point de vue sur le lexique qu'il décrit.

Dans le même état d'esprit, le panel « Dispositifs » propose une représentation schématique de l'ensemble des tables d'un même domaine, notamment pour y créer les liens d'héritage. Dans ce panel et dans celui des « Topiques », il est possible d'attribuer des couleurs aux tables. L'utilisation de ces couleurs est liée à la présentation des résultats d'analyse et sera détaillée dans la section 8. Une extension en cours de réalisation permet d'avoir un aperçu plus complet sur plusieurs dispositifs (figure 4) selon un format vectoriel permettant de zoomer sur les représentations. La figure montre comment un zoom arrière rend compte de la structure globale d'héritage sur les tables de deux dispositifs distincts, ainsi que la manière dont deux couleurs dominantes (rouge et vert) ont été employées pour distinguer les lexiques des deux domaines correspondants.

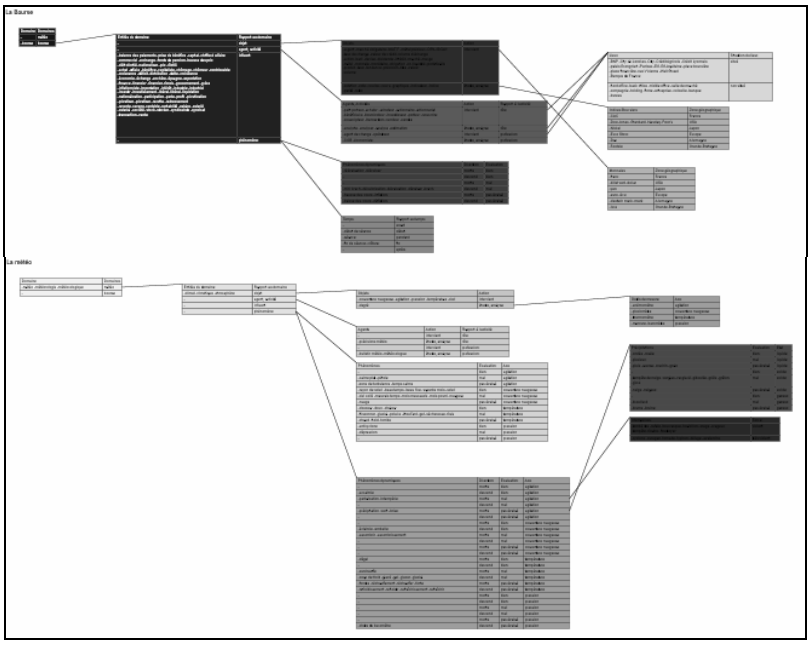


Figure 4 : Dispositifs au format SVG

Les technologies XML (XSLT) permettent de transformer les ressources pour en proposer différentes représentations. Ici, une représentation SVG (langage XML pour les graphiques vectoriels) permet de donner un aperçu de 2 dispositifs. Une couleur a été affectée à chaque table, avec une dominante rouge pour celles du dispositif « La Bourse » (en haut), et une dominante verte pour celles de « La Météo » (en bas).

D'autres applications graphiques sont à l'étude afin de faciliter plus encore la prise en main par différents types d'utilisateurs du modèle et des ressources qu'il permet de décrire. Nous cherchons ici à mettre en évidence la nécessité de combiner, dès la phase de constitution des ressources, une description telle que le modèle la pose et une représentation telle que les interactions l'imposent. Dans notre cas, cela se traduit par une fusion entre structuration lexicale et propriétés de représentations graphiques. La section suivante présente deux applications qui illustrent ce principe.

8. INTERACTIONS POUR L'AIDE À L'INTERPRÉTATION

Comme il a été vu en 4 lors de la présentation du modèle LUCIA, dans les différentes applications envisagées, l'objectif principal est de fournir une aide à l'interprétation. Ceci se traduit dans les faits par la définition d'un modèle interactif plutôt qu'un outil entièrement automatisé. Dans les deux applications que nous présentons ici, le projet ISOMETA et la recherche documentaire, il nous semble possible de distinguer, parmi les besoins d'interaction, des aspects génériques et des aspects envisagés comme

spécifiques, et dont nous étudions la dépendance tant vis-à-vis du modèle que de la tâche et de l'utilisateur.

9. GÉNÉRICITÉ ET CADRES APPLICATIFS

Dans le projet ISOMETA, présenté plus en détails dans Perlerin *et al.* (2002) et Beust *et al.* (2003), nous nous intéressons à la métaphore conceptuelle de la *Météorologie boursière* faisant intervenir deux domaines lexicaux : le domaine source de *la Météorologie* et le domaine cible de *la Bourse*. Cette analyse est menée sur un corpus constitué d'environ 600 articles de Bourse issus du journal *Le Monde sur CDROM* et fait suite aux travaux sur la métaphore présentés dans Ferrari (1997) et Ferrari *et al.* (2000). Le préalable à l'étude de ce phénomène à l'aide du modèle LUCIA est la constitution des dispositifs pour les deux domaines envisagés. Leur apparence globale est celle présentée dans la figure 4, deux couleurs dominantes ayant été associées aux deux domaines : *la Bourse* en rouge, *la Météorologie* en vert. L'analyse automatique consiste essentiellement à détecter les lexies support de redondances d'attributs. L'observation des résultats nécessite de parcourir la collection à la recherche des documents les plus susceptibles de receler des emplois de la métaphore analysée. Pour faciliter ce repérage, nous avons créé une interface (figure 5) regroupant l'ensemble des représentations des documents traités par les modules d'analyse.

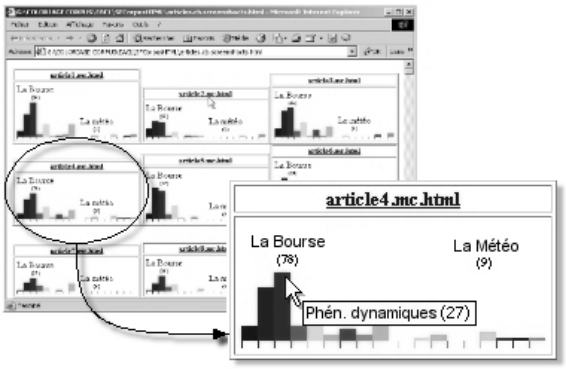


Figure 5 : Visualisation du corpus analysé pour le projet IsoMeta
Une même page HTML regroupe l'ensemble des représentations en histogrammes des documents. Le document « article4 » contient 78 lexies du domaine de *la Bourse* et 9 du domaine de *la Météorologie*.

Dans le projet d'aide à la recherche documentaire, les dispositifs sont utilisés pour filtrer et réordonner des résultats provenant de systèmes classiques tels que les moteurs de recherche de l'Internet. Le préalable est la constitution d'un ou plusieurs dispositifs relatifs aux domaines de la recherche. L'analyse automatique consiste à détecter les redondances d'attributs à différentes échelles, de manière à permettre le repérage d'une

partie de document pertinente, dans un ensemble qui l'est peut-être moins. L'observation des résultats nécessite à nouveau de parcourir la collection à la recherche des documents les plus pertinents pour la recherche envisagée.

Dans les deux applications, le modèle n'est pas en mesure de décider pour l'utilisateur de la pertinence des documents. Il permet juste de les ordonner selon la nature, le nombre et la densité des redondances d'attributs trouvées, et de présenter ce premier résultat. Le parcours de l'ensemble ainsi constitué pour y sélectionner les documents pertinents est une tâche interactive dont la généricité repose sur l'utilisation d'un corpus, d'une collection de documents, pour les deux applications étudiées. Pour cette tâche commune, la généricité est cependant réduite. La disposition de l'ensemble des documents, les fonctionnalités de navigation dans cet ensemble, incluant la possibilité de sélectionner pour l'observation détaillée d'un document particulier, sont autant d'aspects génériques. Différentes solutions existent (Johnson et Schneiderman, 1991 ; Robertson *et al.*, 1991 ou encore Hearst,1995), nous les discutons plus en détail dans Perlerin et Ferrari (2004). La représentation d'un élément dans l'ensemble ne participe plus de la généricité de l'interaction. Chaque document doit être représenté pour permettre à l'utilisateur d'apprécier rapidement sa pertinence, tant absolue que relative à la collection. Nous détaillons en 10 les représentations choisies pour les deux applications, en analysant leurs spécificités.

Dans les deux applications, la phase de navigation est suivie d'une phase de lecture d'un document. Qu'il s'agisse de repérer les emplois métaphoriques ou les zones du document en rapport avec le thème de la recherche effectuée, cette phase présente elle aussi des aspects génériques et des aspects spécifiques. La généricité concerne ici les fonctionnalités de navigation et de lecture pour un document, qui se retrouvent dans de nombreuses applications de TAL et pour lesquelles nous n'apportons pas de solution nouvelle ici. La spécificité des interactions à mettre en œuvre dans cette phase concerne la présentation des résultats d'analyse. Pour le projet ISOMETA, présenter le document pour en permettre la lecture n'est pas suffisant en soi, les emplois métaphoriques doivent être rapidement repérés, et donc mis en évidence au sein du document. En ce qui concerne la recherche documentaire, repérer les zones pertinentes est aussi un besoin, mais ces zones n'ont *a priori* pas la même portée que celles qui sont intéressantes pour les emplois métaphoriques. La présentation des résultats est donc dépendante de la tâche. Nous détaillons dans la suite celle que nous avons retenue pour chaque application étudiée, en continuant d'en analyser la dépendance vis-à-vis d'autres facteurs.

10. DÉPENDANCES VIS-À-VIS DU MODÈLE, DE L'UTILISATEUR ET DE LA TÂCHE

Le modèle sur lequel se fonde une application de TAL constitue un premier facteur de dépendance potentielle pour les interactions et les représentations visuelles. Dans notre étude de cas, ce facteur est fixe et ne peut être modifié. Pour les deux applications, il est possible d'afficher les résultats d'analyse conjointement au texte d'origine ; une partie de ces résultats est superposée au document affiché pour la lecture. Cette possibilité n'est pas systématique, elle tient en partie à l'existence d'analyses locales, qu'il devient donc possible de situer dans le document. D'autres approches, globales, synthétiques, produisent des résultats qui ne peuvent être mis en relation avec un élément particulier du texte. Il convient donc de noter que le modèle influence les moyens d'interaction, même si la présente étude n'en permet pas l'analyse.

L'utilisateur constitue un second facteur de dépendance que la dimension « centrée utilisateur » de notre approche rend prégnant. En effet, c'est à lui que revient la phase finale d'interprétation du matériau textuel. Les deux applications étudiées ont l'avantage de faire intervenir des utilisateurs dont les niveaux d'expertise de la langue, du modèle et des outils qui le mettent en œuvre peuvent être très différents. Or, le niveau d'expertise de l'utilisateur est en rapport direct avec la manière dont les résultats d'analyse sont à présenter. Dans notre approche, son influence commence dès la constitution des ressources lexicales et se poursuit jusqu'à la phase finale d'observation de résultats.

Le projet ISOMETA est destiné à des utilisateurs experts à la fois de la langue, qui est leur objet d'étude, du modèle et des outils. Pour l'analyse d'un fait de langue comme la métaphore, les informations à leur proposer sont nombreuses et complexes, et les représentations visuelles que nous avons élaborées reflètent cette dimension. Pour la phase de navigation dans la collection, la représentation visuelle d'un document analysé consiste en un diagramme interactif (figure 5). Un graphique de type histogramme reflète la distribution des ressources lexicales dans le document analysé. Chaque barre de l'histogramme correspond à une table de même couleur dans un des deux dispositifs, et sa hauteur est proportionnelle au nombre d'occurrence des lexies de cette table trouvées dans le document. L'objectif est de permettre de saisir en un regard la présence de lexique du domaine source de la métaphore étudiée. Lorsque des barres de la couleur dominante du domaine source sont repérées (le vert pour *la Météorologie*), une observation plus poussée de l'histogramme permet d'évaluer plus finement le lexique employé dans le document : le passage de la souris sur l'une des barres déclenche l'affichage du nom de la table associée et du nombre de lexies décrites dans cette table et trouvées dans le document (figure 5). Le

diagramme contient aussi un rappel des noms des dispositifs et le nombre de lexies trouvées pour chacun d'entre eux, ainsi qu'un lien vers le document analysé.

Les Saints de glace BOURSE DE PARIS SEMAINE DU 11 AU 15 MAI
DESSOT ANDRE
Secteur rédaction : Economie Analytique
Sujets en France : BOURSE
le thermomètre a chu
Outils de mesure
Axe : température

Figure 6 : Document traité dans le cadre d'IsoMeta

Visualisation d'un document colorié pour IsoMeta chacune des lexies appartenant aux dispositifs y est mise en valeur par une coloration en arrière-plan avec la couleur de la table correspondante.

Lorsque l'utilisateur décide d'ouvrir le document, il a besoin d'y repérer les emplois métaphoriques potentiels. Nous exploitons encore la couleur pour le guider en surlignant les lexies des domaines décrits par les ressources, comme le montre la figure 6. L'utilisateur peut donc repérer facilement les unités lexicales intéressantes pour sa tâche et observer localement le phénomène étudié. Le passage de la souris sur les unités surlignées déclenche ici aussi l'affichage d'informations complémentaires pour aider à l'interprétation des résultats et éventuellement à réviser les ressources lexicales en conséquence. L'interface actuelle ne rend pas compte des redondances d'attributs. Une interface complémentaire est en cours de développement. Elle exploite un affichage 3D afin de cumuler de nouvelles informations visuelles sur les unités lexicales pertinentes.

Le projet d'aide à la recherche documentaire est quant à lui destiné à tous types d'utilisateurs, en particulier des utilisateurs novices. Pour faciliter la navigation dans les listes de résultats de leur recherche, nous proposons une représentation schématique des documents intégrant une coloration des parties de texte correspondant aux thèmes attendus. Cette représentation, au format SVG (figure 7), permet l'insertion de liens hypertextes pointant directement vers les parties intéressantes des documents. Elle peut être complétée par d'autres informations concernant la langue principale du document, la taille du fichier correspondant, l'URL à laquelle il a été trouvé, etc., comme cela se fait habituellement sur les moteurs de recherche

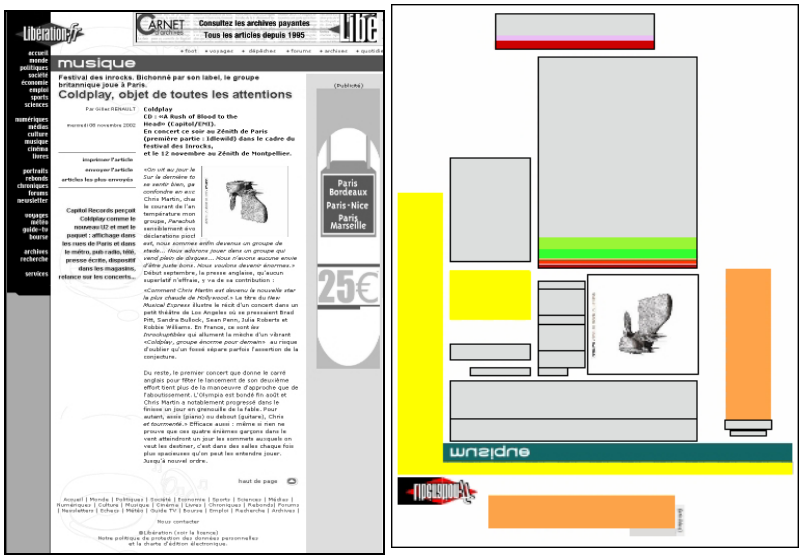


Figure 7 : Visualisation d'un document pour la recherche documentaire
À gauche : un article du journal Libération (format HTML). À droite : la représentation schématique SVG produite automatiquement avec coloriage de certaines parties du texte. Cette représentation peut être réduite et insérée dans une page HTML avec celles des autres documents de la liste envisagée.

Dans la représentation graphique, il est important de conserver l'aspect visuel général du document d'origine afin que l'utilisateur puisse par exemple reconnaître les sources dont il est familier. En revanche, il ne nous a pas semblé primordial de projeter à ce stade autant d'informations que pour le projet ISOMETA. Ce point est cependant à moduler. La spécificité des représentations graphiques des documents peut en effet être adaptée à la manière dont l'utilisateur perçoit et réalise sa tâche. Il est probable qu'un expert effectuant une veille sur un domaine précis propose des ressources détaillées et attende un retour comparable. La même remarque s'applique à la phase finale de lecture du document. Il apparaît finalement que, pour une tâche aussi ouverte que l'aide à la recherche documentaire, où le niveau d'expertise de l'utilisateur n'est pas fixé *a priori*, les interactions et les représentations visuelles à lui proposer ne le sont pas non plus et doivent rester adaptables. Nous présentons dans Perlerin et Ferrari (2004) des travaux issus du domaine de l'IHM (Interaction Homme-Machine) qui pourraient enrichir considérablement les interfaces que nous utilisons sur ces aspects (Card, 1999 ; Jacquemin et Jardino, 2002 ; Lamping, 1995 et Mackinlay et Robertson, 1993).

Les applications étudiées tendent à montrer que ce n'est pas tant la tâche qui guide le choix des interactions et des représentations visuelles, mais plutôt la manière selon laquelle l'utilisateur peut appréhender cette tâche. Lorsque l'utilisateur est supposé expert au départ, toute la richesse du

modèle peut se transposer aux interactions et aux représentations qui lui sont proposées. Mais lorsqu'il n'est familier ni du modèle ni des outils, les interactions et les représentations visuelles utilisées doivent avoir autant l'objectif de l'aider dans sa tâche que celui de le former. Il est donc indispensable d'offrir à l'utilisateur la possibilité de s'approprier pleinement toutes ces notions pour, à terme, améliorer son efficacité dans la réalisation de la tâche concernée. Le modèle LUCIA prévoyant un retour sur les ressources à l'issue d'un cycle d'expérience, la familiarisation avec le modèle passe aussi par un lien fixe entre les ressources et la manière dont il peut y être fait référence lors de la présentation de résultats d'analyse. C'est pourquoi l'association de couleurs aux tables est faite de manière permanente et non pas uniquement lors de chaque phase d'analyse. Prévoir la manière dont il sera fait référence aux ressources lors de la présentation (interactive) des résultats à l'utilisateur, utiliser dès leur constitution des méthodes similaires pour présenter ces ressources améliore selon nous la lisibilité des modèles et leur prise en main. Les interactions et les représentations visuelles véhiculent une grande partie de l'information que la machine communique à l'utilisateur, leur cohérence avec les notions des modèles mis en œuvre aide donc l'utilisateur à se les approprier.

11. CONCLUSION, PERSPECTIVES

L'étude présentée dans cet article repose sur l'analyse d'applications fondées sur un même modèle de TAL. Ce modèle, LUCIA, s'articule autour de l'utilisateur, en lui offrant la possibilité de spécifier ses propres ressources lexicales afin de lui fournir une aide à l'interprétation de documents textuels dans le cadre de différentes tâches. La constitution des ressources a été présentée comme un moyen de familiariser l'utilisateur avec le modèle et les notions sous-jacentes. De ce fait, pour sa mise en œuvre, l'utilisation de représentations proches du modèle nous a semblé préférable à celles qu'offrent les outils informatiques standards. En particulier, les systèmes de gestion de bases de données et la nébuleuse technologique XML ne nous ont pas semblé apporter de réponses satisfaisantes lorsqu'il s'agissait de remplir les fonctionnalités classiques de création et modification des ressources structurées. Une interface spécifique au modèle est selon nous nécessaire, non seulement pour que l'utilisateur puisse s'approprier les notions de ce modèle, mais aussi pour qu'il puisse le faire sans compétence informatique particulière, et sans confondre à l'arrivée un standard de représentation informatique avec un modèle de TAL ! Nous allons donc dans le sens d'une collaboration pluridisciplinaire pour l'élaboration d'outils informatiques adaptés à chaque modèle, notamment lorsque ceux-ci sont destinés à des non informaticiens, plutôt que la réutilisation systématique d'outils standard.

Pour les applications étudiées dans la suite de l'article, bien qu'elles soient en nombre restreint, les besoins d'interactions que nous avons pu dégager, ainsi que leur caractère spécifique ou générique, se retrouvent dans de nombreux travaux de linguistique de corpus. En effet, les phases génériques de navigation dans une collection et de lecture d'un document apparaissent comme peu dépendantes du modèle, de la tâche et de l'utilisateur. Elles sont plutôt directement liées à l'utilisation d'outils de TAL sur une collection. Différents types d'interactions sont envisageables pour réaliser ces deux phases et une réflexion transversale avec le domaine de l'IHM (Interaction Homme-Machine) s'avère nécessaire pour compléter nos travaux. La représentation graphique associée à un document lors de la navigation dans une collection et la superposition d'informations lors de la visualisation des documents sont des interactions qui présentent quant à elles un caractère spécifique indéniable. Elles sont liées au modèle et dépendent selon notre analyse essentiellement de l'utilisateur et de son niveau d'expertise du modèle. Dans le cas particulier du modèle LUCIA, des attributs graphiques sont associés aux ressources dès leur constitution, de manière à permettre une appréhension aisée tant des représentations graphiques que du modèle auquel elles sont ainsi liées.

Pour poursuivre nos travaux, il nous paraît intéressant de mener une étude à plus grande échelle pour mieux caractériser la dépendance des interactions dans les applications de TAL et de linguistique de corpus, en particulier vis-à-vis des modèles linguistiques utilisés. La réflexion sur l'instrumentation informatique de la linguistique doit, selon nous, attacher une part importante aux interactions et aux moyens de visualisations des informations. Il nous semble donc pertinent de revisiter les propositions de standards et plates-formes d'ingénierie linguistiques sous cet angle, en analysant notamment la possibilité d'y intégrer les aspects interactionnels que nous pensons délaissés à l'heure actuelle.

12. RÉFÉRENCES

Beust P. 1998. *Contribution à un modèle interactionniste du sens*, Thèse de doctorat en informatique de l'Université de Caen.

Beust P. ; Ferrari S. et Perlerin V. 2003. *NLP model and tools for detecting and interpreting metaphors in domain-specific corpora*, In Proceedings of Corpus Linguistics 2003. UCREL, Lancaster, p. 114-123.

Card S. K. 1999. *Information Visualization :* Using Vision to Think. Morgan Kaufmann Publishers, p. 1-34.

Coursil J. 1992. *Grammaire analytique du français contemporain - Essai d'intelligence artificielle et de linguistique générale*, Thèse de doctorat en informatique de l'Université de Caen.

Ferrari S. 1997. *Méthode et outils informatiques pour le traitement des métaphores dans les documents écrits*. Thèse de doctorat en informatique de l'Université Paris XI, Orsay.

Ferrari S. ; Giguet E. ; Lucas N. et Vergne J. 2000. “Projet LINGUIX, recherche d’informations et traitements linguistiques : le cas des métaphores.” *In* : *Le document électronique dynamique, Actes du 3ème Colloque International sur le Document Électronique*, CIDE 2000. Lyon, 4-6 juillet 2000, p. 279-293.

Hearst M. A. 1995. “TileBars : Visualization of Term Distribution Information in Full Text Information Access.” In *Proceedings of the Conference on Human Factors in Computing Systems* CHI’95. ACM Press.

Jacquemin C. et Jardino M. 2002. “Une interface 3D multi-échelle pour la visualisation”, In : *Actes d’IHM’2002.*

Johnson B. et Schneiderman B. 1991. “Tree-maps : A space-filling approach to the visualization of hierarchical information structures.” In : *Proceedings of IEEE Visualization ’91*, p. 284-291.

Lamping J. 1995. “A focus + context technique based on hyperbolic geometry for viewing large hierarchies.” In : *Proceedings of the Conference on Human Factors in Computing Systems CHI’95*. ACM Press.

Mackinlay J. D. et Robertson G. G. 1993. “The Document Lens.” In : *Proceedings of the ACM User Interface and Software Technology conference UIST’93*, p. 173-179.

Perlerin V. 2002. “Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes.” In : *Actes de RECITAL 2002.* Tome 1, p. 507-516.

Perlerin V. ; Ferrari S. et Beust P. 2002. “Métaphores et dynamique sémique.” In : *Actes des 2e Journées de la Linguistique de Corpus.* Lorient, septembre 2002.

Perlerin V. et Ferrari S. 2004. “Modèle sémantique et interactions pour l’analyse de documents”, In : *Approches Sémantiques du Document Électronique*, *Actes du septième Colloque International sur le Document Électronique, CIDE.7*. La Rochelle, France, 22- 25 juin 2004, p. 231-251.

Rastier F. 1987. *Sémantique interprétative*. Paris : Presses Universitaires de France.

Robertson G. G.; Mackinlay J. D. et Card S. K. 1991. “Cone Trees : Animated 3D Visualizations of Hierarchical Information.” In : *Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI’91*. New York : ACM Press, p. 189-194.

13. NOTES

(1) source : http://www.portalpes.com/meteo_alpes/page_droite/aNWA.htm

(2) source : *Le Monde sur CDROM*, 1987 à 1989.